DOCTOR OF PHILOSOPHY

Human protein-protein interaction prediction

McDowall, Mark

Award date:
2011

DOCTOR OF PHILOSOPHY

Human protein-protein interaction prediction

Mark McDowall

2011

University of Dundee

HUMAN PROTEIN-PROTEIN INTERACTION PREDICTION

By

Mark McDowall

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT

UNIVERSITY OF DUNDEE

DUNDEE, UNITED KINGDOM

JULY 2011

Contents

List of Tables

List of Figures

Acknowledgements

Professor Geoffrey J. Barton as my supervisor. Michelle S. Scott who started the project and has been a fantastic support throughout my time. Dr Tom Walsh for all of the technical support. Dr Chris Cole for all of the help with Jpred and analysis of the predictions. Dr Jim Proctor for help with all things Java and web services related. The BBSRC for funding my studentship. To everyone in the Barton Group for their friendship and help throughout my time in Dundee.

To my Teeny Lady, Sîan

. . .

Will you marry me?

. . . to Mum and John

Nan and Granddad

UNIVERSITY OF DUNDEE

COLLEGE OF LIFE SCIENCES

I certify that Mark McDowall has satisfied all the terms and conditions of the relevant Ordinance and Regulations to qualify in submitting this thesis in application for the degree of Doctor of Philosophy.

Dated: July 2011

Research Supervisor: _____

Prof Geoffrey J. Barton

UNIVERSITY OF DUNDEE

Date: **July 2011**

Author: **Mark McDowall**

Title: **Human Protein-Protein Interaction Prediction**

Department: **College of Life Sciences**

Degree: **Ph.D.**

I hereby declare that the work described in this thesis is my own; that I am the author of this thesis; that it has not previously been put forward in submission for any other degree or qualification; and that I have consulted references herein.

Mark McDowall

Abstract

Protein-protein interactions are essential for the survival of all living cells, allowing for processes such as cell signalling, metabolism and cell division to occur. Yet in humans there are only $<$ 40k annotated interactions of an interactome estimated to range between 150k to 600k interactions and out of a potential 300M protein pairs. Experimental methods to define the human interactome generate high quality results, but are expensive and slow. Computational methods play an important role to fill the gap.

To further this goal, the prediction of human protein-protein interactions was investigated by the development of new predictive modules and the analysis of diverse datasets within the framework of the previously established PIPs protein-protein interaction predictor (Scott and Barton, 2007). New features considered include the semantic similarity of Gene Ontology annotating terms, clustering of interaction networks, primary sequences and gene co-expression. Integrating the new features in a naïve Bayesian manner as part of the PIPs 2 predictor resulted in two sets of predictions. With a conservative threshold, the union of both sets is $>$ 300k predicted human interactions with an intersect of $>$ 94k interactions, of which a subset have been experimentally validated.

The PIPs 2 predictor is also capable of making predictions in organisms that have no annotated interactions. This is achieved by training the PIPs 2 predictor based on a set of evidence and annotated interactions in another organism resulting in a ranking of protein pairs in the original organism of interest. Such an approach allows for predictions to be made across the whole proteome of poorly characterised organisms, rather than being limited only to proteins with known orthologues.

The work described here has increased the coverage of the human interactome and introduced a method to predict interactions in organisms that have previously had limited or no annotated interactions. The thesis aims to provide a stepping stone towards the completion of the human interactome and a way of predicting interactions in organisms that have been less well studied, but are often clinically relevant.

Chapter 1

Literature Review

Preface

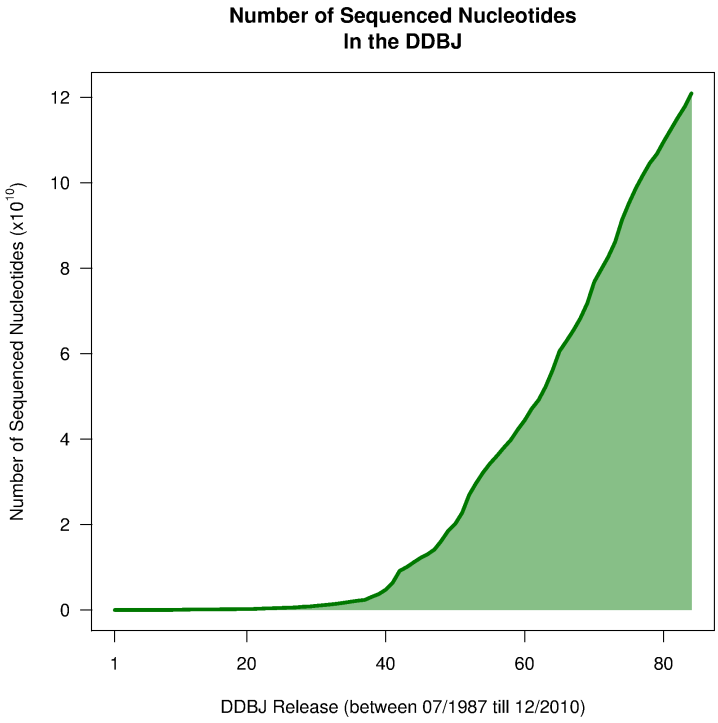
This Chapter introduces the prediction and validation of protein-protein interactions. This Chapter also explains Bayesian statistics and their application within the Protein-Protein Interaction Predictor, PIPs.

1.1 Introduction

Biology is going through an age of rapid growth in the volumes of data that are being generated via high throughput analysis. One area of exponential growth is in high throughput genome sequencing. Figure 1.1 shows the rate of growth of the number of sequenced nucleotides within the DDBJ sequence database (Kaminuma et al., 2011). By the end of 2010, there were more than 4000 fully sequenced and publicly available genomes (Cochrane et al., 2011). In addition to sequencing further high throughput methods, such as microarrays, that enrich the information about

a genome and an organism. Challenges brought about by these methods include making sense of the large volumes of information and interpreting the data in a biologically meaningful way.

Figure 1.1: Number of sequenced nucleotides present in the DDBJ (Kaminuma et al., 2011) (Date 23-03-2011).



One way in which the data can be analysed is by assigning function to proteins within the genome. This is possible via the transfer of annotation based on sequence homology to a protein of known function, but as the number of sequences grows, these methods become harder to apply to more evolutionary divergent species (Rost et al., 2003). One way of compensating for this is to use other sources of information, such as structure, phylogeny, genomic data and protein-protein interactions (Sleator and Walsh, 2010). A source of information that can help in the annotation of a protein's biological function are protein-protein interaction networks. These

interaction networks do not rely on sequence homology between proteins to infer function, but on the interaction of pairs or sets of proteins.

Protein-protein interactions (PPI) allow the cell to perform and regulate key processes. Protein-protein interactions can be either transient or stable and are often part of a larger protein complex. Transient interactions, such as enzymatic interactions or cell signalling, are short term interactions. Stable interactions such as those involved in complex formation often provide a functional unit, such as the nuclear pore or the proteasome.

The purpose of identifying protein-protein interactions is to understand the functional and dynamic properties of the cell (Stelzl and Wanker, 2006). Many groups have started to identify the protein-protein interactions within whole proteomes (Uetz et al., 2000; Ito et al., 2001; Giot et al., 2003; Li, 2004; Lehner and Fraser, 2004; Formstecher et al., 2005; Rual et al., 2005; Stelzl et al., 2005; Arifuzzaman et al., 2006). There are many experimental and computational methods that aim to identify protein-protein interactions. Methods such as X-ray crystallography, mass spectroscopy, and biochemical/biophysical experiments, can identify binary or complex interactions related to a protein of interest. Some methods have been scaled up for high throughput analysis to identify protein-protein interactions over a whole proteome.

1.2 Experimental Determination of Protein-Protein Interactions

Table 1.1 summarises 15 experimental methods for identifying protein-protein interactions. Many of the methods were developed for the analysis of binary or single complex interactions, some have been scaled for use in a high throughput manner. Description of the methods has been divided up into their most common application in modern research.

1.2.1 Low Throughput Methods

Until recently, the investigation of protein-protein interactions has mainly been done using low throughput methods. X-ray crystallography is the most accurate way to determine the structure and interaction between two proteins. There are $\geq 13,000$ structures listed in the Protein Data Bank (PDB) (24^{th} March 2011) that have been determined via X-ray crystallography with resolutions ranging between 0.5Å to 6Å, however the majority of these are the same protein with different ligands bound, after filtering at 90% sequence homology, that number drops to just ≥ 3000 structures.

With NMR it is possible to determine the 3D structure of proteins and protein complexes in solution (Bonvin et al., 2005) and also resolve transient interactions. There are ≥ 2000 human protein structures in the PDB (06^{th} July 2011) that have been resolved via NMR.

Method	**Throughput**	**Type of Interaction**	**Reference**
X-Ray Crystallography	L	B/C	Robinson et al. (2007)
Nuclear Magnetic Resonance	L	C	Bonvin et al. (2005)
Cryo Electron Microscopy	L	C	Rossmann et al. (2005)
Pull-Down Assays	L	B/C	Singh and Asano (2007)
Co-Immunoprecipitation	L	B	Singh and Asano (2007)
Surface Plasmon Resonance	L	B	Lofas and Johnsson (1990); Jost et al. (1991)
Fluorescence Resonance Energy Transfer	L	B	Yan and Marriott (2003); Jameson et al. (2003)
Fluorescence Correlation Spectroscopy	L	B/C	Yan and Marriott (2003); Müller et al. (2003)
Atomic Force Microscopy	L	B	Yang et al. (2003)
Yeast-2-Hybrid	H	B	Fields and Song (1989); Chien et al. (1991)
Tagged Affinity Purification	H	C	Rigaut et al. (1999); Puig et al. (2001)
Protein Microarray	H	C	Zhu et al. (2001)
Gene Co-Expression	H	I	Schena et al. (1995, 1996)
Synthetic Lethality	H	I	Tong et al. (2001); Simons et al. (2001b)
Flow Cytometry	H	I	Sklar et al. (2007)

Table 1.1: Methods of protein-protein interaction determination. Throughput method: High throughput, H; Low throughput, L. Type of Interaction: Binary, B; Complex, C; Inferred via further analysis, I

In cryo electron microscopy (cryo-EM) an array of protein complexes (10^3 to 10^5 protein complexes) are frozen in a variety of orientations and then imaged at a resolution of between 7Å to 10Å (Rossmann et al., 2005). From cryo-EM images it is possible to reconstruct a 3D model of the complex (Xiao and Rossmann, 2007). Cryo-EM allows very large complexes to be imaged at low resolutions, for example Lander et al. (2006) used cryo-EM and X-ray crystallography to resolve the structure of the p22 Viron with a resolution of 17Å.

Pull-down assays are a useful method for analysis of protein-protein interactions *in vitro*. An example method is to express a bait protein fused with Glutathione S-transferase (GST). Cells expressing the bait protein are then lysed and bait proteins are selected for with glutathione linked resin. In binding the bait protein it also pulls out any binding partners (the prey). The prey can then be separated with SDS-PAGE and the prey proteins identified by mass spectroscopy (Singh and Asano, 2007). Whole cell extract can be used when there are no known interactions, but confirmation of the interactions must be performed with a purified version of each of the prey protein extracts to establish a true interaction between bait and prey proteins.

Co-immunoprecipitation is a pull down assay for protein-protein interaction identification. When the cell is lysed the bait protein is selected for using an antibody that is attached to a Protein-A or Protein-G bound resin. The bait protein along with any associated proteins (the prey) is precipitated out by centrifugation (Singh and Asano, 2007).

Surface plasmon resonance (SPR) is a method based on interaction between a photon and the electrons of a thin metal film (usually Gold (Lofas and Johnsson,

1990)). The angle of reflection of the incident light changes proportionally to the change in mass concentration in the vicinity of the metal surface (Jost et al., 1991). Changes to the local mass concentration, such as a protein attached to the metal film going from a non-interacting state to an interacting state therefore increases the local mass concentration, this changes the angle of reflection of the incident light. Proteins are located near the surface of the metal by fixing them in a methyl-modified dextran hydrogel meaning that SPR is a label-free detection method, therefore there is no possibility of the tag modifying the folding of the protein (Lofas and Johnsson, 1990).

There are also methods of directly analysing the interaction of a protein complex. These methods include fluorescent techniques such as Fluorescence resonance energy transfer (FRET) (Yan and Marriott, 2003; Jameson et al., 2003) or Fluorescence correlation spectroscopy (FCS) (Yan and Marriott, 2003; Müller et al., 2003). Atomic Force Microscopy (AFM) can directly probe two interacting proteins with electron microscope (Yang et al., 2003).

1.2.2 High Throughput Methods

With the size of proteomes ranging from thousands to hundreds of thousands of proteins (when splice variants are taken into consideration); various methods have been developed to identify true protein interactions on a large scale (Figure 1.2).

Yeast Two-Hybrid, Y2H

The Y2H method is an *in vivo* technique to investigate binary interactions within the cell (Figure 1.2a) (Shoemaker and Panchenko, 2007). It uses an eukaryotic

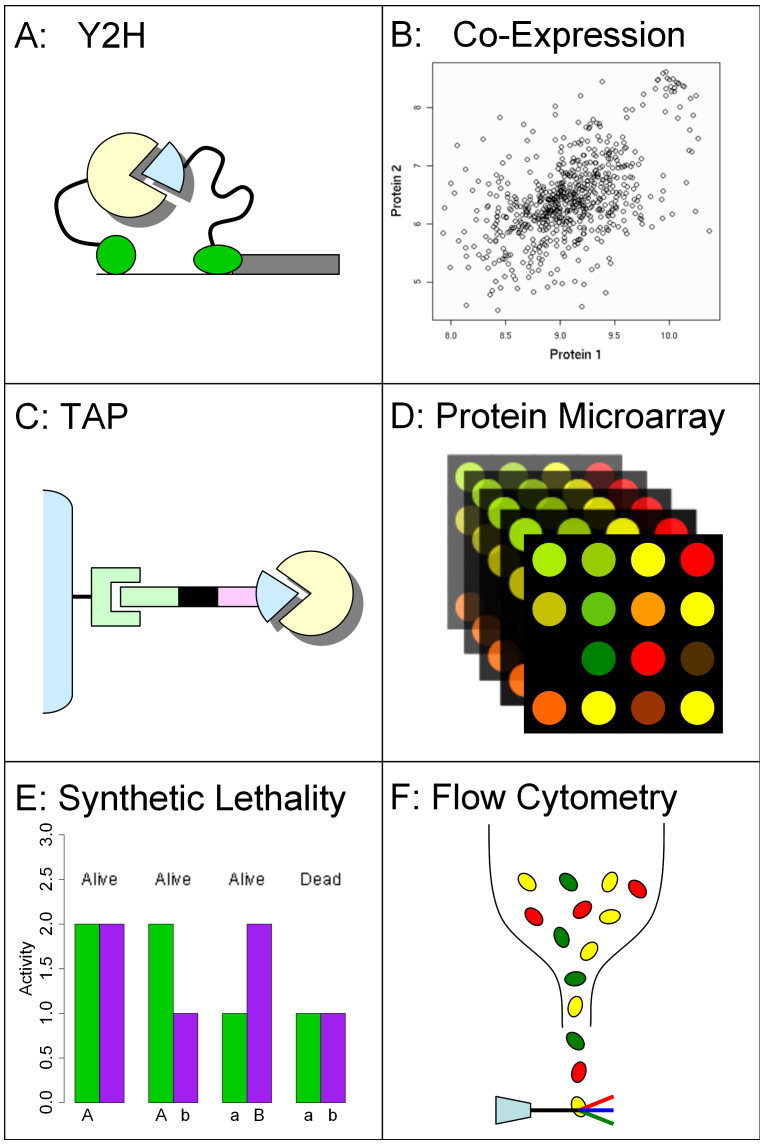


Figure 1.2: Diagrammatic representation of experimental techniques to determine PPI; adapted from Figure 1 (Shoemaker and Panchenko, 2007). A, Y2H detects interaction between protein X and protein Y, where X is linked to a binding domain and Y is linked to the activation domain of a reporter gene; B, Gene co-expression analysis, dark areas show genes whose expression is highly correlated; C, TAP analysis allows the extraction of whole complexes via an IgG-binding domain (green), Tobacco ETCH Virus (TEV) cleavage site (black) and a calmodulin binding protein (red); D, Protein microarray detects binary PPIs, immobilised proteins on a solid phase are probed by tagged proteins; E, Synthetic lethality is used to determine if proteins have a similar function within the cell where mutation of 1 of the proteins is non lethal, but mutation of both is lethal; F, Flow Cytometry could be used to determine the presence of an interaction between two proteins and the temporal phase of the cell cycle.

transcription factor where the two domains, a DNA binding domain and transcript activation domain, have been separated. When separated, activation of the reporter gene is lost. To regain activation of the transcription factor, the two domains need to be located close enough to interact. Locating the two domains on separate target proteins allows for the activation of a reporter gene if the two target proteins interact. The first implementation of this method used the yeast transcription factor GAL4 and the reporter gene LacZ (Fields and Song, 1989; Chien et al., 1991).

The first proteome to be analysed by Y2H was the *Escherichia coli* bacteriophage T7 (Bartel et al., 1996). The T7 bacteriophage was selected due to a small genome (39,937bp) coding for only 55 proteins. Many high throughput Y2H studies have been performed in diverse organisms including *E. coli* (Arifuzzaman et al., 2006), worm (Li, 2004), fly (Giot et al., 2003; Formstecher et al., 2005), yeast (Uetz et al., 2000; Ito et al., 2001) and a broad spectrum study of the human proteome (Rual et al., 2005; Ramani et al., 2005). There are also other high throughput human Y2H experiments that have focused on specific subsets of proteins, such as Lim et al. (2006) which focuses on 54 proteins involved in ataxia, but they ended up discovering 770 novel interactions.

There are two variations of Y2H used in genome analysis:

Library: Libraries of genes are fused with the GAL4 activation domain and libraries of ORFS (Open Reading Frames) fused with the DNA binding domain with a plasmid. Successful transformant cells containing the plasmids with the activation domain are pooled and hybridised with each of the individual DNA-binding transformants. When an interaction is found after selecting for

transformants that have both plasmids, the proteins are then determined via DNA sequencing (Bartel et al., 1996; Uetz et al., 2000; Ito et al., 2001).

Matrix Array: Similar to the library method, but each of the DNA-binding domain set are hybridised with the transformants from the activation domain library separately (Uetz et al., 2000).

It was found that over the same *S. cerevisiae* dataset the matrix array method generated an average of 3.3 potential protein-protein interactions per protein, compared to 1.8 potential interactions per protein for the library method (Uetz et al., 2000). Even though the library method does not provide as many potential interactions as the array method, it was found to be easier to scale for high throughput analysis (Uetz et al., 2000).

Despite the great utility of this approach, Y2H suffers from several drawbacks. Uetz et al. (2000) found that some proteins could initiate transcription by themselves thus polluting the inferred interactions. Also, the addition of the extra binding and activation domains can result in the protein not folding correctly thus inhibiting interaction. Certain types of proteins, such as membrane proteins, are not amenable to this type of study without extensive modification of the method. As a result there is a high risk of false positive and false negative interactions (Chien et al., 1991). The accuracy of this method was initially estimated to be 50% when analysing a yeast interactome (Sprinzak et al., 2003). Therefore additional methods are required to validate the results produced by Y2H studies. The problems of Y2H were acknowledged by Stanley Fields who later remarked that had they focused on these problems during development, they may never have developed the technique (Fields,

2009).

Tandem Affinity Purification (TAP) and Mass Spectroscopy (MS)

Designed as a pull down assay, TAP has been scaled for high throughput analysis to identify binary and complex interactions. TAP is a method to extract whole protein complexes from a cell by using a TAP cassette connected to a bait protein (Figure 1.2c). The TAP cassette consists of the IgG-binding domain from the *Staphylococcus aureus* Protein A and the calmodulin binding protein separated by a spacer containing the Tobacco ETCH Virus (TEV) cleavage site. The TAP cassette is fused to a bait protein by the calmodulin binding domain. The bait protein construct is inserted into the host cell to be expressed. To extract the bait protein and the associated protein complex the cells are lysed and then the cell extract is filtered using beads coated with IgG that bind to the IgG binding domain in the TAP cassette. After flushing the beads to remove excess cell extract it leaves the bait protein bound to the beads. The protein complex is cleaved from the beads using the TEV protease. To extract the complexes from the protease, beads coated with calmodulin are used to filter out the complexes, which are then washed and the complexes released from the beads using EGTA (ethylene glycol tetra-acetic acid) ready for analysis using MS (Rigaut et al., 1999; Puig et al., 2001).

MS is an effective method for the identification of protein sequences (Borch et al., 2005). Depending on the way a protein is ionised there are two main types of ionisation used for MS to analyse protein-protein interactions; matrix-assisted laser desorption/ionization (MALDI) and Electron Spray Ionisation (ESI) mass spectroscopy (Aebersold and Mann, 2003). With large databases of peptide fragments it

is possible to reconstruct the protein based on the mass spectrum of a protein. TAP allows the automation of MS to identify the proteins of different protein complexes in succession.

TAP allows for the verification of protein-protein interactions with high accuracy, although the tag can act to partially or fully block the binding site. Due to the multiple washing procedures it makes TAP a poor choice for identifying transient protein-protein interactions.

Gene Co-Expression

Proteins that interact, especially those in stable interactions, tend to be co-expressed, ensuring all interactions are present in appropriate amounts. Gene co-expression can therefore be used to infer protein-protein interactions. A method for analysing co-expression of genes was developed in 1995 with the genome of *Arabidopsis thaliana* because it is a higher eukaryotic organism with one of the smallest genomes (Schena et al., 1995). A year later Schena et al. (1996) described the expression patterns of 1046 human cDNAs.

In gene expression analysis known DNA sequences are heat treated and spot printed to a glass slide to form a DNA chip. Each spot represents a single gene. mRNA extracted from the cell is reverse transcribed to form cDNA that is tagged with a fluorescent probe. The cDNA is hybridised with the DNA sequences on the DNA chip. When the samples are exposed to a laser the intensity of the emitted radiation is proportional to the level of expression (Schena et al., 1995) (Figure 1.2b). By comparing multiple cell cultures from different tissue types it is possible to determine which genes are co-expressed. Even though gene co-expression analysis

can take a systematic view of the level of gene expression in an average cell the method is limited to only being able to infer interaction and further work is required to verify whether an interaction has occurred.

Protein Microarray

In a protein microarray proteins are bound to the solid phase (Figure 1.2d). A protein library array is spot printed to the surface of a glass slide via an exposed lysine residue creating a Schiff's base linkage (MacBeath and Schreiber, 2000), these are the bait proteins. A prey protein is then washed over the slide allowing it to interact with the bait proteins. The prey proteins are tagged with a fluorescent marker highlighting which bait proteins it interacts with. Yeast was the first whole eukaryotic proteome to be analysed using protein microarrays (Zhu et al., 2001). However, the downside of this method is that the markers can act to block the binding site of the protein, inhibiting the identification of an interaction.

Synthetic Lethality

Synthetic lethality can determine if proteins within the cell function in parallel (Figure 1.2e). In a synthetic lethal interaction, the deletion of one gene does not affect the cell, but the deletion of both genes is fatal (Tucker and Fields, 2003). This technique had been applied to yeast (Tong et al., 2001) and human (Simons et al., 2001a,b) tissue cultures for high-throughput analysis. Synthetic lethality allows the robustness of the biological network to be studied in an organism and allows for the inference of protein complex membership (Ye et al., 2005) and the analysis of protein-protein interaction at the functional level. However this method would not identify two functionally linked proteins if there is a redundant pathway present.

Flow Cytometry Analysis

Flow cytometry is a method for the analysis of cells suspended in solution. It can analyse cells at a rate of 50,000 per second. By tagging the proteins, mRNA transcripts or DNA, it is possible to determine the level of co expression, concentration or the phase of the cell cycle (Sklar et al., 2007). By measuring the scattering of emitted fluorescence from the cell it is possible to determine various properties about that cell. Forward scatter of photons infers the size of the cell, whereas the side scatter of photons is a function of the granularity of the cell (Bonetta, 2005). The advantage that flow cytometry has over other techniques is that the cell does not have to be lysed to determine the content, plus each cell can be measured independently rather than homogenising a culture of cells that could be in many different stages of the cell cycle. Datasets of individual cells with knowledge about co expression and cell cycle phases would allow for a temporal aspect to be applied to the analysis of PPIs (Figure 1.2f).

Currently, there are no publicly available datasets as this is still a relatively new concept to combine cytomics and proteomics. With groups starting to work on methods for applying flow cytometry to the proteome (Bernas et al., 2006) and the creation of the Human Cytome Project (Valet, 2005) this technique will provide a unique dataset for the creation of a protein interactome.

1.3 Computational Prediction of Protein-Protein Interactions

Computational methods can be used to complement experimental methods. The benefit that computational methods provide is that they are cheaper and quicker, they can incorporate information that has been derived from multiple experimental sources and they can be applied to abstract methods of analysis, such as sequence analysis or literature mining. For the identification of interacting pairs of proteins, computational methods can highlight the most likely pairs of proteins to guide experimental studies and render them more cost effective. Computational methods also have the capability of covering the whole proteome whereas only a few experimental methods allow this in practice.

1.3.1 Prediction Methods

Supervised predictive methods aim to infer the interaction of two proteins based on known examples. The focus of the predictive method can range from predicting the structural conformation of two proteins and whether they interact (protein docking), to predicting the probability of protein-protein interactions over a whole proteome based on information about the whole genome.

There are many protein structural models in public repositories, such as the PDB

(http://www.rcsb.org/pdb/), but the experimental methods for producing structural models of complexes are slow. Computational methods, such as protein docking, have been developed to use the known protein structures to predict protein-protein interaction. In 2001 CAPRI (Critical Assessment of PRedicted Interactions) was set up as a community experiment to measure the accuracy of modelling protein-protein interactions.

Protein-protein interaction prediction using protein models initially uses rigid-body search algorithms that search the protein surface for optimal binding sites using a fast Fourier transformation method (FFT). The predicted binding sites are then analysed for the most optimal interaction by scoring each protein-protein interaction by considering properties such as residue-residue interaction, electrostatics and hydrogen bonding (Smith and Sternberg, 2002). In CAPRI, protein-protein interaction prediction now also focuses on the optimisation of scoring functions (Lensink et al., 2007). Groups focusing on protein-protein interaction prediction submitted 1994 models, of which only 5.1% were of acceptable to medium quality. Groups focusing on scoring optimisation who applied their algorithm to the submitted models were able to identify 31.7% of the acceptable to medium quality predicted models (Lensink et al., 2007).

The first methods of computational protein-protein interaction prediction for a whole genome were developed for the newly sequenced prokaryotic genomes. Computational methods of prediction included:

Gene Fusion: A gene fusion event is when a gene in an ancestral genome has become separated into two transcribed genes in a descendant genome. To

maintain the function of the ancestral gene product it is hypothesised that the two descendant gene products must interact. Gene fusion can also work the other way around, where two ancestral genes have become fused over time (Enright et al., 1999; Marcotte et al., 1999). This method of predicting interaction has been implemented as part of the STRING (von Mering et al., 2007) and CODA (Reid et al., 2010).

Co-Localisation: Co-localisation refers to the conservation of gene order within a bacterial genome. The hypothesis is that the products of genes that are close to, or neighbouring, each other on the genome, are more likely to interact than the products of genes whose locations are greatly separated (Dandekar et al., 1999). It was found that some genes in eukaryotes, which are involved in the same biological pathway and would interact, tended to be transcribed on polycistronic mRNA (Blumenthal, 1998), for example GDF1 and UOG identified by Lee (1991).

Co-Occurrence: Co-occurrence describes the simultaneous presence and evolution of pairs of genes in a correlated fashion (Pellegrini et al., 1999). It is hypothesised that this co-occurrence infers a functional relationship and hence an increased chance that the two gene products will interact. The PLEX webserver was developed to implement phylogenetic profiling to predict functionally linked proteins (Date and Marcotte, 2005).

Later methods were able to take advantage of the numerous sequenced genomes to predict protein-protein interaction via orthology (Matthews et al., 2001; Lehner and Fraser, 2004). The hypothesis is that if two proteins are found to interact in

the genome of one species, if the genome of a second species has a homologous pair of genes, the products of those genes are likely to interact (Matthews et al., 2001). Recent studies by Ramani et al. (2008) used co-expression patterns in human and the expression of orthologous genes from other organisms to predict protein-protein interactions in humans. Ramani et al. (2008) was able to predict the 7000 protein-protein interactions of which 1411 were known interactions and 5589 predicted novel protein-protein interactions. Based on cross validation the accuracy of to predictor was calculated to be $54 \pm 10\%$ (Ramani et al., 2008).

Sequence based methods of prediction have been applied to search for commonly occurring patterns between interacting proteins and differentiating them from non-interacting proteins. Chinnasamy et al. (2006) used hydrophobicity to represent the protein sequence and was able to predict protein-protein interactions in yeast with a reported accuracy of 80% to 84%. Shen et al. (2007) predicted protein-protein interactions in humans with a reduced residue alphabet with an accuracy of $> 82.23\%$.

Another sequence based method is the identification of "hot spots". These are regions of the surface of a protein that are critical for the interaction of two proteins and are often detected via alanine-scanning mutagenesis (Cunningham and Wells, 1989). Prediction of these residues often involves considering the structural data of a complex, such as Kortemme and Baker (2002) and web services like HSPred (Lise et al., 2011). Unlike Ofran and Rost (2007b) who use a sequence similarity approach based on the ISIS server (Ofran and Rost, 2007a) allowing for the prediction of interaction sites on a protein without the need for a 3D structure.

Protein disorder is believed to promote protein-protein interaction (Tompa and

Fuxreiter, 2008). It is also believed that disorder allows for regulation between active and inactive states of a protein via protein modification, or ligand binding, and that there is a careful balance between the ordered and disordered state (Zhang et al., 2007). Attempts have been made to include the proportion of disorder within the prediction of protein-protein interactions (Scott and Barton, 2007), although it has been suggested that the level of disorder could be proportional to the number of potential interactions a protein may have (Hegyi et al., 2007).

It is also possible to use text mining as a way to extract pairs of proteins from a paper and classify them by whether they interact or not. A simplistic method is the prediction of interaction based on two genes being present in the same abstract/article (Marcotte et al., 2001; von Mering et al., 2005; He et al., 2009). More sophisticated methods employ natural language processing methods to analyse the context of the genes within a sentence or paragraph (Kim et al., 2008). However, one of the first problems with these methods is being able to recognise gene/protein names, with rules or dictionaries there is still the potential for a high false-positive rate (Jang et al., 2006). Using natural language processing Jang et al. (2006) was able to achieve a precision of 83% at identifying protein-protein interactions from abstracts present in PubMed.

There is a wide range of accuracies calculated by different predictive methods, however Hart et al. (2006) estimated the false positive rates for high throughput experimental and computational methods for yeast and humans to be 72% and 90% respectively. The method that Hart et al. (2006) used to estimate the false positive rate was proposed by D'haeseleer and Church (2004) by comparing the protein-protein interactions between different interaction assay sets (Lehner and Fraser,

2004; Rhodes et al., 2005; Stelzl et al., 2005; Rual et al., 2005) and a reference dataset that was derived from the Human Protein Reference Database (HPRD) (Mishra et al., 2006), Biomolecular Interaction Network Database (BIND) (Alfarano et al., 2005), Reactome (Joshi-Tope et al., 2005; Ramani et al., 2005).

1.3.2 Machine Learning Methods For The Prediction of Protein-Protein Interaction

There are numerous machine learning algorithms that can be applied to predict whether a pair of proteins interact given a set of measurements. The most basic form is to define a way of measuring a difference between the values assigned to two proteins and then setting a threshold over which the proteins are predicted to interact and below which they are predicted to not interact. The training method can take measurements derived from information such as experimental data (see Section 1.2) or those described above in Section 1.3.1 and then use this information for the classification of a protein pair. The machine learning methods allow for the determination of the optimal thresholds for the classification of interaction dependent on a given set of evidence for a set of training examples, the selection of which is discussed in Section 1.5.

The most common machine learning methods that have been applied to the prediction of protein-protein interactions include:

- Artificial Neural Networks
- Support Vector Machines (SVMs)
- Bayesian Classification (see Section 1.6)

Artificial neural networks (ANNs) are based on the concept of a perceptron that is able to take a set of inputs and calculate a regression of the input data and then provide a binary output as the result. The multilayered ANN has a set of nodes (perceptrons) that are linked via connections that weight the outputs of each of the nodes, the weights are modified most commonly via a back-propagation method proposed by Rumelhart et al. (1986). ANNs have a wide range of applications, from facial recognition (Mitchell, 1997) to modelling of biological networks and from modelling networks of neurones in the brain to the prediction of protein structures (Cole et al., 2008). Ofran and Rost (2007b) used ANNs to identify hot spot binding sites on proteins by using the networks to integrate numerous sources of information about the protein, from amino acid composition to homology. Neural networks have also been used to identify the interacting surfaces of proteins based on the primary and 3D structure using protein complexes present in the PDB (Zhou and Shan, 2001; Fariselli et al., 2002; Wang et al., 2010). Zhou and Shan (2001) considered sequence profiles and spatial neighbouring of residues, while Fariselli et al. (2002) and Wang et al. (2010) also included evolutionary conservation. All three methods obtained accuracies between 69% to 73%.

Support Vector Machines (SVMs) have been widely used for the prediction of protein-protein interactions (Ben-Hur and Noble, 2005; Shen et al., 2007; Deng et al., 2009; Lin et al., 2010) as well as the prediction of interaction interfaces (Wang et al., 2006). The SVM algorithm was originally designed by Vapnik in 1963 as a linear classifier, but it was not until 1992 that it was possible to generate non-linear classifiers with the inception of the kernel trick (Bolser et al., 2003). The Kernel trick allowed for the mapping of real values in a multidimensional space

for the resolution of a linear regression of the data. The benefit of SVMs comes from being able to handle large numbers of input values and to derive an optimal regression of real data for classification. Often SVMs are implemented to represent the strings of residues within a protein, either as motifs, proportion of different amino acids or values to represent physico-chemical properties of the proteins, such as hydrophobicity or disorder. The classifier designed by Deng et al. (2009) was trained on an ensemble of information about a protein's structure and amino acid sequence, including PDB structures, multiple sequence alignments, accessible surface area and sequence profiles. The training set was derived from two previous studies (Conte et al., 1999; Chakrabarti and Janin, 2002) where the complexes had been obtained from the PDB, but they were not species specific. Unlike Shen et al. (2007) who trained SVMs on protein sequence profiles and obtained precision values of $\geq 82\%$ and sensitivities of $\geq 84\%$. SVMs can also be used to integrate many distinct sources of information, such as the Predicted Arabidopsis Interactome Resource (PAIR) (Lin et al., 2010) and Ben-Hur and Noble (2005) in yeast, but the accuracy of the two methods varies widely, Lin et al. (2010) cite a sensitivity of 48% in comparison to Ben-Hur and Noble (2005) when the specificity is 99% with a sensitivity of 80%.

In the study by Wang et al. (2006) to predict protein interaction interfaces, they used the same dataset as the Wang et al. (2010) study (69 non-redundant protein complexes derived from the set used by Fariselli et al. (2002)). The SVM was built based on sequence profiles and the evolutionary rate, which resulted in a predictor with an accuracy of 65% when identifying regions of a protein sequence that were likely to be part of a protein-protein interaction.

Bayesian classification, described in greater detail in Section 1.6.3, has in the

past 10 years become more widely used for the prediction of protein-protein interactions on a proteome scale. Bayesian methods allow for the integration of multiple sources of evidence to infer the likelihood of interaction between a pair of proteins. Bayesian methods have used a diverse range of sources to infer interaction from orthology (Lee et al., 2008) or sequence (Chinnasamy et al., 2006; Burger and van Nimwegen, 2008), whereas other predictors incorporate mixtures of other sources of information, such as gene co-expression, micro-array, Gene Ontology annotations, sequence information (Jansen et al., 2003; Rhodes et al., 2005; Scott and Barton, 2007). Jansen et al. (2003) developed their predictor within yeast and using a likelihood ratio threshold of 600 they found that they were able to predict that a pair of proteins had a 50% chance of being part of the same complex. Rhodes et al. (2005) and Scott and Barton (2007) developed naïve Bayesian predictors in human, however with similar likelihood ratio thresholds, 381 and 400 respectively, they obtain false positive rates of 50% (Rhodes et al., 2005) and 76% (Scott and Barton, 2007). When compared using the methods described by D'haeseleer and Church (2004) and Hart et al. (2006) to the new interactions present in the October 2006 release of the HPRD the false positive rate for predictions made by Rhodes et al. (2005) increases to 78%. However, these false positive rates are below the average false positive rate for high throughput experimental data, 90% (Hart et al., 2006).

The advantage of ANNs and SVMs is the ease of use and that they require fewer training examples to generate capable predictors. However, they can be rather opaque methods when it comes to determining the strength of each source of evidence in the prediction of the final classification for the identification of protein-protein interactions. This is an important advantage of the Bayesian methods, it

is possible to determine the contribution of each piece of information that is available for predicting whether two proteins are likely to interact or not. However, the disadvantage for the Bayesian methods is that they require larger training datasets, which is often difficult, especially for protein-protein interaction prediction. With the increase in the number of available protein-protein interactions that can be used for the training of a Bayesian predictor it has meant that they are now a viable option. Therefore the selection of which tool to use is dependent on the availability of examples for training and the availability of predictive data.

1.4 Protein-Protein Interaction Databases

Table 1.2 summarises 13 protein-protein interaction databases that provide experimentally or computationally predicted interactions. Many of the databases have interactions that have been derived from multiple species and multiple experimental types (high and low throughput, see Section 1.2). The HPRD (Peri et al., 2003; Mishra et al., 2006) is a source of high quality human protein-protein interactions that have been derived from the literature. The majority of the interactions within the HPRD have at least one piece of evidence based on low throughput experiments with only 2% coming from evidence only based on Y2H studies.

The Database of Interacting Proteins (DIP) (Salwinski et al., 2004), IntAct (Kerrien et al., 2007a), Mammalian Protein-Protein Interaction database (MPPIs) (Pagel et al., 2005) and Reactome (Vastrik et al., 2007) are all curated sources of protein-protein interactions. IntAct and MPPIs have been extracted from the literature and manual curation, although MPPIs is focused on mouse interactions whereas IntAct

is species independent. DIP is a collection of protein-protein interactions that have been derived from high and low experimental data with the high throughput data being assigned a reliability score (Salwinski et al., 2004). Reactome is different from DIP, IntAct and MPPIs as it is concerned with interactions that represent biological pathways within the cell (Pagel et al., 2005).

Database	Description	No. Proteins	No. Interactions	Reference
HPRD	CL (2% H)	25661	38167	(Peri et al., 2003; Mishra et al., 2006)
IntAct	CLH	15000	23586	(Kerrien et al., 2007a)
MPPIs	CL	460	N/A	(Pagel et al., 2005)
Reactome	CLH	2499	N/A	(Vastrik et al., 2007)
DIP	CLH	1224	1794	(Salwinski et al., 2004)
BioGrid	LH	6374	30761	(Stark et al., 2006)
MINT	LH	6106	20832	(Chatr-aryamontri et al., 2007)
SNAPPIdb	S	N/A	5677 (Domains)	(Jefferson et al., 2007)
iPfam	S	N/A	N/A	(Finn et al., 2005)
OPHID	P	N/A	47221	(Brown and Jurisica, 2005)
PIPs	P	22889	37606	(Scott and Barton, 2007; McDowall et al., 2009)
POINT	P	N/A	38151	(Huang et al., 2004)
STRING	LHP	1513782 (373 species)	N/A	(von Mering et al., 2003)
Bacteriome	LH	N/A	4863 (*E coli*)	(Su et al., 2008)

Table 1.2: Human Protein-Protein Interaction Databases, unless stated. Data derived from: L = Low throughput; H = High throughput; C = Curated database; P = Predicted interactions; S = Structural data. * STRING uses interactions imported from (Mishra et al., 2006; Vastrik et al., 2007; Salwinski et al., 2004; Stark et al., 2006; Chatr-aryamontri et al., 2007; Alfarano et al., 2005; Kanehisa et al., 2004)

SNAPPIdb (Jefferson et al., 2007) and iPfam (Finn et al., 2005) are databases of three dimensional protein domain-domain interactions, although iPfam is based only on Pfam domains; SNAPPIdb draws annotations from SCOP, CATH and Pfam, thus increasing the number of non-redundant domain-domain interaction annotations. SNAPPIdb also classifies the interactions by the orientation of the domains involved in the interaction.

Three databases are listed in Table 1.2 that have predicted human protein-protein interactions derived from orthologous data; OPHID (Brown and Jurisica, 2005), PIPs (Scott and Barton, 2007; McDowall et al., 2009) and POINT (Huang et al., 2004) all use orthologous data from other species that have more complete interactomes. OPHID and PIPs also use information that has come from experimental data, such as co-expression and gene annotations.

1.5 Dataset Selection

Most computational predictors of protein-protein interactions are based on supervised learning methods, which identify predictive characteristics that differentiate protein pairs that do or do not interact. Supervised machine learning methods therefore require sets of training examples to represent the distribution, or patterns, of the system to make accurate prediction and for which the final performance of a classifier can be assessed (Mitchell, 1997). Often two datasets are required for training, a positive and a negative dataset. The positive dataset represents known information about the system. For example with protein-protein interaction prediction, this dataset would represent known pairs of proteins that interact within a

given species. The negative dataset, given the previous example, would represent pairs of proteins that are known not to interact.

When it comes to protein-protein interactions, Gold Standard Positive (GSP) and Negative (GSN) datasets for training are typically sets of known positive and negative interactions that have been experimentally confirmed. In the cell, there is a huge bias in the size of the number of positive interactions in ratio to the number of protein pairs that do not interact. The datasets should therefore also be able to represent this imbalance (Jansen and Gerstein, 2004).

1.5.1 Positive Datasets

Positive datasets are usually sourced from publically available databases of protein-protein interactions (see Section 1.4). There are numerous interaction databases that can be used to generate positive datasets, including BioGRID (Stark et al., 2006), DIP (Salwinski et al., 2004), HPRD (Keshava Prasad et al., 2009), IntAct (Aragues et al., 2006; Kerrien et al., 2007a), MINT (Ceol et al., 2010) and MPPIs (Pagel et al., 2005). The positive set can then be selected based on the interactions present in one or many databases. However, which database(s) is selected can depend on a multitude of factors, such as number of available interactions, method of interaction identification, species dependence and even previous experience. For example, databases such as the DIP, HPRD, IntAct and MPPI are curated databases of protein-protein interactions and are derived from low and/or high throughput experimental data; whereas the protein-protein interactions in BioGRID and MINT are derived from low and/or high throughput experimental data, but not curated. In contrast, databases such as the HPRD are literature curated databases where

the interactions are identified by reading the literature and annotating proteins that interact.

The use of the databases listed above is not without its problems. Often there is little overlap between databases and there are errors within the databases, both from human curation of the data and from misinterpretation of the results (Cusick et al., 2009). This can lead to gold standard positive datasets that can contain pairs of proteins that do not interact. Some predictors, such as Support Vector Machines (SVMs) (Ben-Hur and Noble, 2006), are more capable of handling misclassifications, whereas others would be more sensitive. The expectation is that over time the quality and number of interactions increases as more people add to the database and edit the entries to remove errors.

1.5.2 Negative Datasets

The selection of negative protein-protein interactions is not as simple as selection of positive interactions. Currently there is only a single database of negative protein-protein interactions, called The Negatome (Smialowski et al., 2009), although IntAct does have a tag within the database (IS:0257) to identify protein pairs that do not interact. This relies on it being used by the biologists submitting the data and currently remains unpopulated. Unfortunately the Negatome contains less than 1000 protein pairs that do not interact for the human proteome, this is a small fraction of the potential number of protein pairs that in reality do not interact. As a result other methods are often used to generate training and testing sets to represent protein pairs that do not interact.

Several protein-protein interaction prediction methods (Jansen et al., 2003; Rhodes

et al., 2005) use randomly selected protein pairs that were annotated to localise to different compartments of the cell for their negative dataset. However, even though this generates good quality negative interaction sets with low numbers of contaminating false negatives it has been found to introduce bias when building predictors where the negative examples have been chosen in this manner (Ben-Hur and Noble, 2006). As a result the performance of the classifier is artificially increased above its real accuracy as the predictor was making predictions based on whether two proteins were co-localised even though not all co-localised protein pairs interact.

The second method commonly used for the generation of the negative dataset is by randomly selecting protein pairs and filtering out known interactions, which provides a less biased view for the training of an interaction predictor (Gomez et al., 2003; Ben-Hur and Noble, 2005; Qi et al., 2006). Even though the bias is removed, it does make it more difficult for the predictor and Ben-Hur and Noble (2006) found that there was a decrease in the accuracy of the predictor.

The problem that plagues both methods of selection is that the complete interactome remains unknown and so in selecting negatives, there is the chance that unknown positive interactions could also be selected. The chance of these false negatives being included in the negative set is increased for the second method as the proteins can be part of the same compartment, which increases significantly their likelihood of interaction. However, because estimates of the size of the complete human interactome range between 130k to 650k interactions (Hart et al., 2006; Stumpf et al., 2008; Venkatesan et al., 2009) and the potential number of all binary interactions ($> 300M$), the chances of selecting two proteins that actually interact and are included in the negative dataset are very small.

A third method has also been proposed to predict negative interactions (Shoyaib et al., 2009). Shoyaib et al. (2009) use a graph based method to identify negative interactions by determining that proteins that are further apart within the network are less likely to interact than those that are closer to each other. However this method still lays prey to the problem of an incomplete interactome.

Therefore, the ideal way to select gold standard positive and negative examples to be used for training will only be known once the whole interactome has been identified, but by that point the construction of a predictor of protein-protein interaction is a moot point.

1.6 Bayesian Classification

There are two main methods for estimation and hypothesis testing, Frequentist and Bayesian. Reverend Thomas Bayes first formalised the Bayes theorem, which came to light when the paper was posthumously published by his friend Reverend Richard Price (Bayes and Price, 1763). Bayes's posthumous paper refers to rolling a ball (W) on a table and then rolling a second ball (O) n times and the number of times that O lands to the right of W are counted (X), where X is the representation of the probability of a binomial success.

It was not until Pierre-Simon Laplace worked on the formalisation of the method and the publication of his work on Inverse Probability originally in 1774 in French, but translated by Stigler in 1986 to English (Stigler, 1986b; Laplace, 1774), later to become known as Bayesian inference (Stigler, 1986a), that Equation 1.6.1 was proposed.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Equation 1.6.1: Bayes' Theorem

Where h is the hypothesis and $P(h)$ is the probability that h is true, this is known as the prior probability. D is the training data and therefore $P(D)$ is the probability of D given no prior knowledge about which hypothesis is true. $P(D|h)$ is the probability of observing D given that h is true. $P(h|D)$ is the posterior probability, which is the probability of the hypothesis being true given the data. The objective of Bayes Theorem is to find the hypothesis with the maximal posterior odds over a given hypothesis space.

The probability $P(D)$ can be calculated by Equation 1.6.2.

$$P(D) = \sum_{i=1}^{n} P(D|h_i)P(h_i)$$

Equation 1.6.2:

Where n is the number of potential hypotheses. This means that Equation 1.6.1 can be rewritten as Equation 1.6.3 to test multiple mutually exclusive hypotheses.

$$P(h_1|D) = \frac{P(D|h_1)P(h_1)}{\sum_{i=1}^{n} P(D|h_i)P(h_i)}$$

Equation 1.6.3:

Equation 1.6.3 calculates the probability that h_1 is true given the training set D. This also allows for multiple hypotheses.

Example There are 2 urns, A and B. Urn A contains 15 white balls and 85 black balls, urn B contains 50 white balls and 50 black balls. If a white ball is picked at random, what is the probability that it was selected from urn A?

If h_1 represents the ball originating from urn A and h_2 from urn B, then $P(h_1)$ and $P(h_2)$ are the probabilities of each hypothesis being true. $P(h_1)$ and $P(h_2)$ are equal because from the pickers point of view the urns are identical therefore $P(h_1) = P(h_2) = 0.5$. D is the observation that a white ball has been selected. Given that $P(D|h_1) = \frac{15}{100}$ and $P(D|h_2) = \frac{50}{100}$ and using Equation 1.6.3:

$$\begin{aligned} P(h_1|D) &= \frac{P(D|h_1)P(h_1)}{P(D|h_1)P(h_1) + P(D|h_2)P(h_2)} \\ &= \frac{0.15 \times 0.5}{0.15 \times 0.5 + 0.5 \times 0.5} \\ &= 0.23 \end{aligned}$$

Therefore given the selection of a white ball, the probability of the ball being selected from urn A is 0.23.

Within Equation 1.6.1 there are 3 terms that are required to calculate a probability, $P(D|h)$ can be easily calculated, $P(h)$ and $P(D)$ are more tricky. $P(h)$ is the prior probability and therefore requires knowledge about the training set or dataset as a whole. $P(D)$ is difficult to calculate as many different hypotheses can give rise to the same value. $P(D)$ can be removed from the equation by comparing the ratio of two hypotheses, h and h'. Given that:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad \text{and} \quad P(h'|D) = \frac{P(D|h')P(h')}{P(D)}$$

The ratio of the two expressions cancels out the probability $P(D)$ to give:

$$\frac{P(h|D)}{P(h'|D)} = \frac{P(D|h)P(h)}{P(D|h')P(h')} = \frac{P(h)}{P(h')} \times \frac{P(D|h)}{P(D|h')}$$

Equation 1.6.4:

Equation 1.6.4 is the likelihood ratio form of the Bayes theorem (Barnard, 1949) where $\frac{P(D|h)}{P(D|h')}$ represents the likelihood ratio. The ratio of $\frac{P(h)}{P(h')}$ is known as the prior odds ratio (O_{prior}), but is still difficult to calculate. In a Bayesian manner the estimation of the prior odds ratio is dependent on prior knowledge about what would be expected at random, usually calculated by sampling. To illustrate the calculation of the O_{prior} for the prediction of likelihood of protein-protein interaction prediction, if there are 33309 interactions between 8968 proteins the O_{prior} would equal $\frac{1}{1206}$ (see Equation 1.6.5 and Chapter 3.2.2).

$$O_{prior} = \frac{\frac{33309}{(8968^2)/2}}{\frac{(((8968^2)/2)-33309)}{((8968^2)/2)}} = \frac{33309}{((8968^2)/2) - 33309} = \frac{1}{1206}$$

Equation 1.6.5: Calculation of the prior odds ratio O_{prior} based on the figures from Chapter 3.2.2.

Subsequently for a given training set of 100 positive and 10000 negative protein-protein interactions. If 60 of the positive pairs and 2 negative pairs are probability of a of having a particular feature (LR = 3000 as calculated by Equation 1.6.6) given that they interact. Therefore if the $O_{prior} = \frac{1}{1206}$ and the likelihood ratio is 3000 then the posterior odds ratio (O_{post}) represented in Equation 1.6.4 as $\frac{P(h|D)}{P(h'|D)}$ would

equal 2.5. The interaction is thus 2.5 times more likely to occur than to not occur.

$$\frac{P(D|h)}{P(D|h')} = \frac{\frac{60}{100}}{\frac{2}{10000}} = 3000$$

Equation 1.6.6: Example of the calculation of the likelihood ratio.

A range of thresholds can be selected and the likelihood ratio calculated for each one. For protein pairs that have a given score, they are assigned the likelihood ratio based on the threshold limits.

1.6.1 Naïve Bayesian Classification

The naïve Bayesian classification is a simplification of Bayesian classification making it a more tractable method of learning. Bayesian classification acts to classify an instance given a set of feature values, ie $P(h_j|a_1, a_2, a_3, \ldots a_n)$. Naïve Bayesian Classification makes the assumption that each feature set is independent and therefore the probability of observing an instance given a set of features is the product of the probabilities (Mitchell, 1997). Equation 1.6.1 can therefore be rewritten as Equation 1.6.7.

$$P(h|a) = \frac{P(a_1, a_2, a_3, \ldots a_n|h)P(h)}{P(a_1, a_2, a_3, \ldots a_n)}$$

Equation 1.6.7: Modification of Bayes Theorem

A Naïve Bayesian classifier is a simplification of Equation 1.6.7, which assumes that each dataset (a) is conditionally independent of the others. Therefore to observe the events $a_1, a_2, a_3, \ldots a_n$ given h is true is equivalent to the product for the individual events given h is true (Equation 1.6.8).

$$P(h|a_1, a_2, a_3, \ldots a_n) = \frac{P(h)\prod_i P(a_i|h)}{P(a_1, a_2, a_3, \ldots a_n)}$$

Equation 1.6.8: Naïve Bayesian Classifier

Equation 1.6.8 presents the same problems as the Bayesian version with regards to knowing the probability of all possible events, so can be compared to the probability of a second hypothesis, $P(h')$ (Equation 1.6.9).

$$\frac{P(h|a_1, a_2, a_3, \ldots a_n)}{P(h'|a_1, a_2, a_3, \ldots a_n)} = \frac{P(h)}{P(h')} \times \frac{\prod_i P(a_i|h)}{\prod_i P(a_i|h')}$$

Equation 1.6.9:

As with Equation 1.6.4, Equation 1.6.9 still poses the problem of calculating the prior odds ratio. However, Equation 1.6.9 allows for the comparison of two or more hypotheses across multiple datasets to derive the most likely prediction based on the evidence.

1.6.2 Bayesian versus Frequentist

The use of Bayesian Inference, however, is not without its critics, of which Ronald Fisher was one of the most vocal and prolific in commenting on Inverse Probability (Bayesian Inference) (Fisher, 1930), oddly this note also includes the earliest reference to "Bayesian Inference".

There are subtle differences to the way that probability is perceived between the two methods when estimating the uncertainty of a given set of data. The Frequentist view is that the probability is the long-running expected frequency of an event ($P(A) = \frac{n}{N}$ where n is the number of times that A occurs in a dataset of

N). The Bayesian view is where the probability is the belief in the event being true given incomplete knowledge.

For a Frequentist to calculate a true mean it would require sampling of data from the whole space. From this it is then possible to say that given a specific interval (eg 95%) the true mean is within the interval. As a result, Frequentists look for repeatability of the data and then assign a measure of significance based on the repeatability of the data.

The problem with the Frequentist view is that it requires accurate calculation of what is the long-run value. The sampling of the data needs to be performed enough times and cover as much of the hypothesis space to ensure that the sampled distributions match that of the real distribution (Cox, 2006).

The Bayesian view on this is from the perspective that the data provided is real, what is the believability of the data to match the real distribution and thus the true mean. This allows a Bayesian to reason that for a given interval there is a 95% probability of it containing the true mean. However, this is only possible if there is assumed prior knowledge about the system.

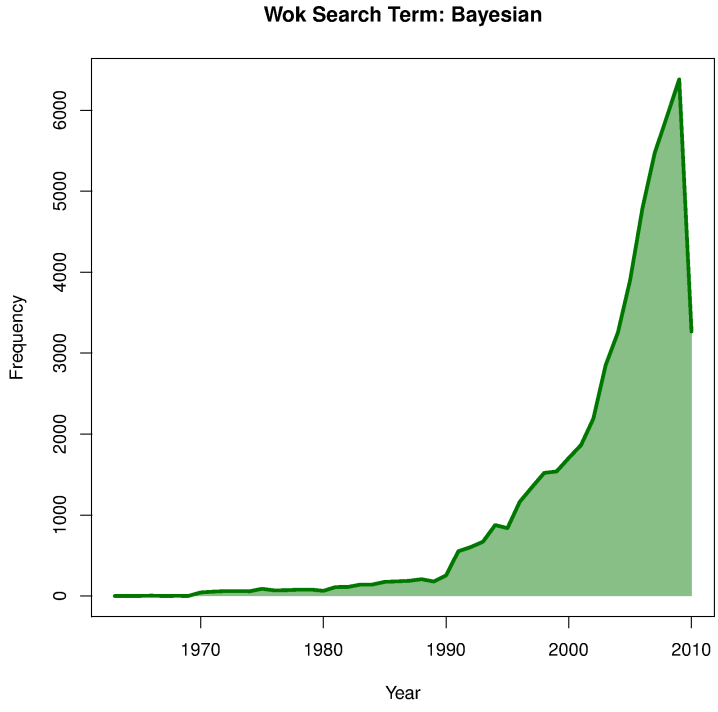
There are two ways to derive a prior probability. The first is to use a prior based on observational data that is not represented in a statistical form, but of a consistent nature and therefore repeatable given similar data. The problem with this is that there needs to be evidence to back up the behaviour of the decisions that are made. The second method is to set a fixed prior probability using a rational degree of belief by taking a Frequentist approach and considering what is learned given the dataset and knowledge about the data (Cox, 2006).

1.6.3 Bayesian Inference and Protein-Protein Interaction Prediction

Figure 1.3 shows the number of publications present in the Web of Knowledge database that include the word Bayesian, this indicated that even though Bayesian inference has been around for over 250 years, there has really only been a wider acceptance and use of the method in the last 20 years. This is in part due to two reasons; the first is down to the acceptance of Bayesian views and the second due to the extra computational power required to test multiple hypotheses and the generation of a justifiable prior probability.

Figure 1.4 shows the number of citations from Figure 1.3 where the articles also include the terms "protein" and "interaction". This is a dramatic reduction from

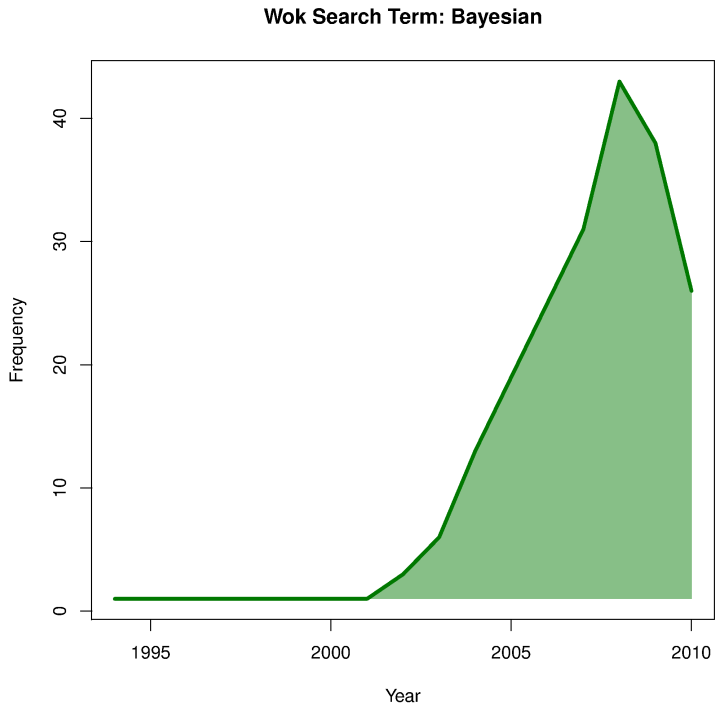
Figure 1.3: Papers that mention "Bayesian" based on a keyword search of Web of Knowledge (10-08-2010).



the 50,000 down to just 209 articles. Of the 209 articles, only 69 also include the term "prediction", so there are still only a few papers that are using Bayesian terms for the prediction of protein-protein interactions.

The use of Bayesian Inference, in particular naïve Bayesian Classification, was shown to be comparable to methods such as decision trees (Mitchell, 1997) for classification (Friedman et al., 1997). Built on the capability of naïve Bayesian classifiers, one of the first attempts at a genome wide application of Bayesian inference was to predict the localisation of proteins within the proteome of yeast (Drawid and Gerstein, 2000). The importance of using holistic data about the cell and its integration for predicting an interactome was highlighted by Mark Gerstein (Gerstein et al., 2002) and one way of doing this would be to use Bayesian Inference. The

Figure 1.4: Papers that mention "Bayesian", "Protein" and "Interaction" based o a keyword search of Web of Knowledge (10-08-2010).



advantage of using naïve Bayesian methods is that it is possible to integrate information from multiple sources. Bringing together information from multiple sources poses the risk of not all evidence sources having evidence for certain interactions (missing data). This can be handled by the naïve Bayesian Classifiers unlike other prediction methods where a value would have to be inferred or those proteins ignored.

One year after the publication of Gerstein et al. (2002), a study was published investigating the prediction of protein-protein interactions in yeast using a naïve Bayesian Classifier (Jansen et al., 2003), based on gene expression, overlap of biological function (Gene Ontology) and essentiality for survival (from MIPs). This highlighted the capability of the naïve Bayesian method to integrate multiple sources of evidence which covered different sets of protein pairs.

With respect to humans the first draft interactome was released in 2004 (Lehner and Fraser, 2004) based on conserved interactions between species to infer interactions in humans. The first method to apply naïve Bayesian inference in humans was by (Rhodes et al., 2005), who used a similar method to (Jansen et al., 2003), but also included orthologous information about interactions in other species. The concept has been taken further to also include analysis of the predicted network of interactions by including a measure of transitive interactions (Scott and Barton, 2007), for full details of the method see Section 1.8 and for developments of the PIPs framework see Chapter 2.

One of the major problems for the prediction of protein-protein interaction, especially in humans, has always been the lack of known protein-protein interactions to use as learning examples. The first protein-protein interaction predictor to use

a naïve Bayesian Classifier was only able to consider 11,678 protein-protein interactions (Rhodes et al., 2005) that were derived from the HPRD (Peri et al., 2004). Now there are over 39,000 binary protein-protein interactions present in the HPRD (Keshava Prasad et al., 2009), but this is still only a fraction of the potential number of protein-protein interactions within the human interactome, which is estimated to be between 120,000 to over 600, 000 (Hart et al., 2006; Stumpf et al., 2007; Venkatesan et al., 2009). Bayesian methods allow the incorporation of knowledge about the potential size of the interactome, along with knowledge about well studied subnetworks. Examples include the setting of the prior odds ratios within predictors (Rhodes et al., 2005; Scott and Barton, 2007).

1.7 Training and Testing Classifiers

During the construction of a classifier it is important to know how predictive it is and that changes that are made are an improvement and not a detriment. To this end there are several methods that can be employed to maximise the chances of increasing the capabilities of a classifier and be able to calculate the accuracy and the performance.

1.7.1 Cross Validation

Cross validation was proposed as a way to divide a given dataset for use in the assessment and comparison of predictive models (Picard and Cook, 1984), but has been utilised since at least the early 1930's (Stone, 1974). It involves taking a dataset, dividing it up into a set number of non-overlapping groups and then leaving one chunk out for testing and training on the rest. There are several forms of cross

validation and the selected method depends on a range of factors, such as the size of the available datasets, avoidance of over-fitting of the training set and available computational power.

Over fitting occurs when the classifier is trained and tested to fit a particular case very well, but when presented with new data it is unable to accurately classify the new test data. Cross validation is able to solve this problem by training on k variations of the data so that potential local optimal classifiers are avoided in preference for a global preference over the training data. Development of a classifier that is more general allows for more accurate predictions when tested on a blind test set (see Section 1.7.5).

There are several variations on cross validation (Mosteller and Tukey, 1977). The first method is to split the dataset in two, one is used for training and the second is used for testing. However, splitting the data requires that you have enough to split the complete dataset in half. The second method is called k-fold cross validation where the dataset is randomly split into k sets, the predictor is trained on k-1 sets and then tested on the final set. K-fold rotates through all potential k combinations and testing on the held out sub set. The k-fold method allows for the predictor to train on and test on all examples within the dataset thus highlighting variance in the performance of the predictor. The third method takes the k-fold to the extreme where k is equal to the number of examples in the dataset, this is also known as Leave one out cross validation (Mosteller and Tukey, 1977).

1.7.2 ROC Curves - Analysis of Performance

Receiver Operating Characteristic (ROC) curves are plots used to assess the performance of a classifier (Fawcett, 2006). Originally developed during World War II to assess the ability to identify enemy planes using radar, they are now used in the wider machine learning and data mining community (Swets et al., 2000).

Some predictors, such as decision trees are only capable of generating a discrete classification. This would relate to a single point within a ROC plot. Other classifiers, such as Bayesian Classifiers are able to assign a proportion, or score, to the classification being correct. Classifiers that are able to assign a continuous value to a classification allow for the production of a ROC curve, whereas a classifier that gives a discrete binary output can only generate a single point.

Given a classifier that assigns a value to the likelihood of a given classification it is possible to generate a ROC curve by varying the threshold for classification over a range of potential values and generating a confusion matrix (Table 1.3) for each threshold point. From the confusion matrix it is possible to calculate several basic statistics.

$$FPR = \frac{FP}{N}$$

Equation 1.7.1: False Positive Rate, FPR

$$TPR = \frac{TP}{P}$$

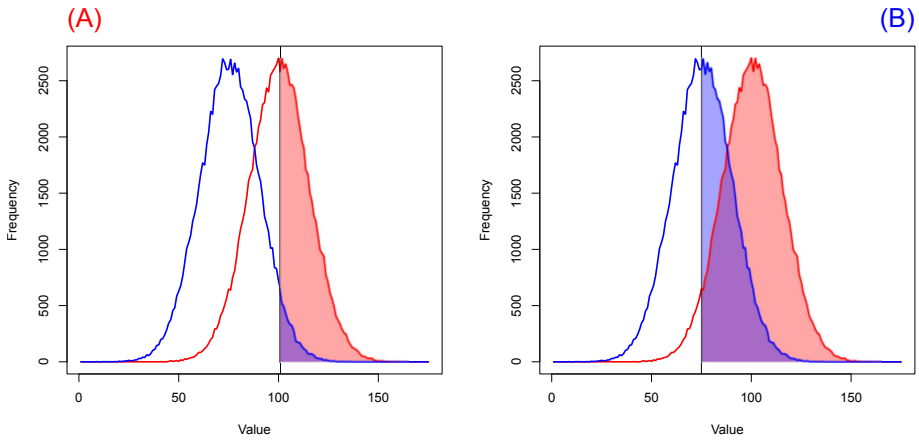
Equation 1.7.2: True Positive Rate, TPR

From two sample sets P (positive) and N (negative) is it possible to plot the

Table 1.3: Confusion Matrix

		Classification	
		p	**n**
Real Class	**P**	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

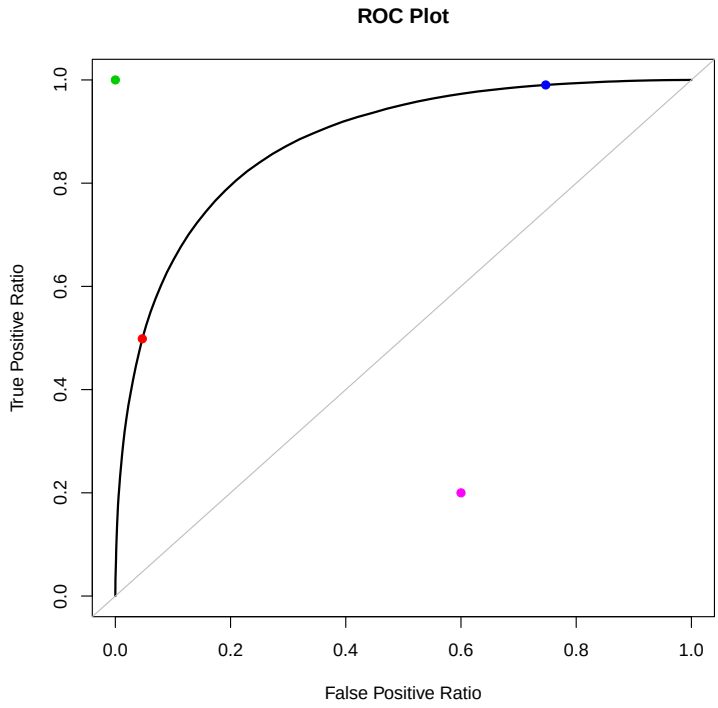
Figure 1.5: Distributions of values for Positive (Red) and Negative (Blue) datasets. The shaded areas represent a positive classification with preselected threshold value. A: the threshold is set at 101; B: the threshold is set at 75.



distribution of scores that are assigned to each example in sets P and N, this is shown in Figure 1.5. As the threshold line in Figure 1.5 is moved, examples to the right are classified as positive (p) and those to the left as negative (n). As the threshold is moved from right to left the proportion of examples that are classified p increases and so the proportion of p:n increases as represented by the shaded areas under the curves in Figure 1.5. The change in proportion of p:n can be visually represented as a ROC plot (Figure 1.6) where the True Positive Rate (TPR, Equation 1.7.2) is plotted against the False Positive Rate (FPR, Equation 1.7.1). Figure 1.6 represents the ROC curve for the classification of the P and N datasets from Figure 1.5.

The plot area in Figure 1.6 represents the space of potential performance points

Figure 1.6: ROC plot based on the distribution of datasets from Figure 1.5. The red dot (●) represents the TP to FP ratio if the threshold is set to 101 (Figure 1.5, panel A); the blue dot (●) represents the TP to FP ratio if the threshold is set to 75 (Figure 1.5, panel B). The green dot (●) represents a perfect prediction where all the positives are classified a positive and all the negatives are classified as negative. The grey line represents the expected curve if classification was random. The magenta dot (●) represents a classifier that performs worse than random.



that a classifier can have. The grey line along the diagonal represents the expected values for a random classifier. The aim is to develop a classifier that is within the upper left triangle of the plot as this represents a classifier that performs better than random. For example the green spot in Figure 1.6 represents a perfect classifier that is able to label all positive example as positive, and all negatives as negative whereas the blue and red classifiers are mostly predictive, but do make some misclassifications. The magenta point represents a classifier that is worse than random, in this case the logic within the classifier can usually be reversed so that the classifier moves into the upper triangle.

The red and blue points on Figure 1.6 represent the two threshold points shown in Figure 1.5, plot A and plot B respectively. The curve represents all the potential threshold points within Figure 1.5. What the ROC curve identifies is that for all threshold points the classifier performs better than a random classifier.

Area Under the ROC Curve (AUC)

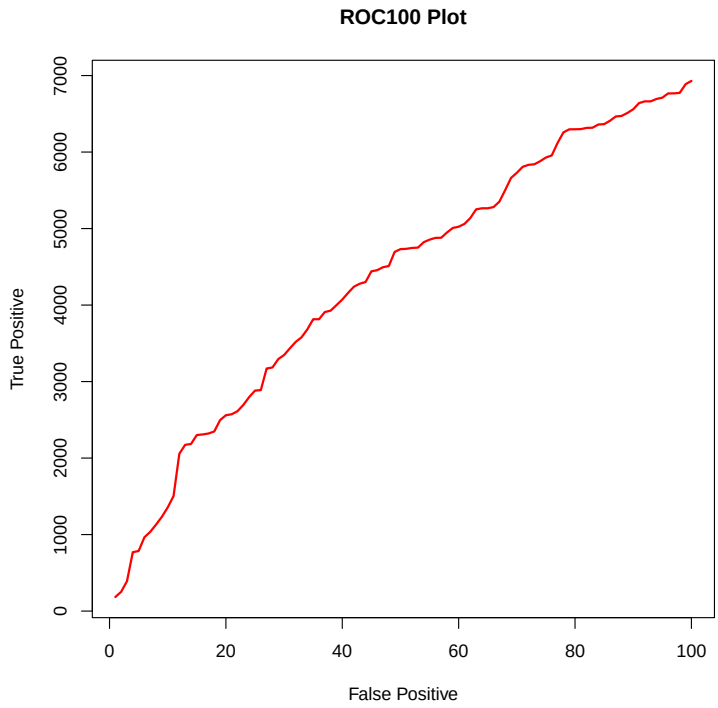
When comparing multiple classification methods in ROC space, calculating the area under the ROC curve is a compact way to assess the performance of each with a single value. The value ranges between 0 and 1, but because random has a value of 0.5 the aim is for each classifier to be > 0.5 and then greater than each other, with the best predictor getting an AUC value of 1.0.

1.7.3 ROCN Curves

One problem that arises with ROC curves is when there is a large negative to positive ratio and correct positive assignment is more important than negative ones. If the training sets are unbalanced with a large number of negative to positive examples it then becomes difficult to distinguish between two classifiers and their profiles over a given proportion of the ROC plot. In this case it is possible to use a ROCN plot where rather than plotting the ratio of p:n, the plot is of the number of true positive predictions that occur over a fixed number of false positive predictions. This allows for the focus to remain on the useful range of the predictor rather than over the whole range of all potential predictions that could be made. False negatives can then be tested for in later releases of the positive training sets.

Figure 1.7 is a ROCN plot for the datasets shown in Figure 1.5. A ROCN

Figure 1.7: ROC100 plot.



is created by ordering the negative examples by their assigned score from highest to lowest. For the top N false positive predictions, the number of true positive predictions with a score greater than or equal to the score of the N^{th} negative example is plotted. These types of plots have been used previously by Ben-Hur and Noble (2005) to access the accuracy of a protein-protein interaction predictor.

1.7.4 Combining Cross Validation and ROC Plots

ROC curves, both ROC and ROCN, are created based on the test set post training. If combined with cross validation it is possible to estimate the mean accuracy of the predictor by averaging over the number of times the predictor is trained and tested during cross validation. This averaging also allows for a greater confidence to be assigned to a predictor especially when comparing between two different methods.

There are two ways to average ROC curves and the selection is dependent on the features of interest (Fawcett, 2006). ROC curves can be averaged vertically for a given false positive rate, or threshold averaged at given scoring threshold points. Each method has its advantages, for example vertical averaging is easy to compute and it provides an error for a given false positive rate. Vertical averaging is appropriate for ROCN plots where the number of false positives is fixed. Threshold averaging allows for an independent variable to act as the fixed point around which the true positive and false positive rates are calculated.

1.7.5 Blind Test Sets

Also known as a Double cross validation (Mosteller and Tukey, 1977) blind test sets are used to test a predictor where the values have been derived independently (temporally or methodologically) from those used for training and parameter estimation. The purpose is to assess the predictive capability of the classifier over unseen data. It is for this reason that it is performed as the last test on the selected final predictor to determine the actual accuracy of the predictor on an unseen dataset.

The selection of a blind set has to be independent of the training and testing dataset during development. In protein-protein interaction prediction a new release of the source database can be used and the accuracy of the predictor can be assessed based on the new interactions. However, it has to be a fair test. For example if a classifier of protein-protein interactions is trained and tested using interactions inferred by genetic interaction, it would then be unfair to use a blind test set containing only interactions for proteins that physically interact due to the differences in the semantics of protein-protein interaction.

1.8 PIPs Framework

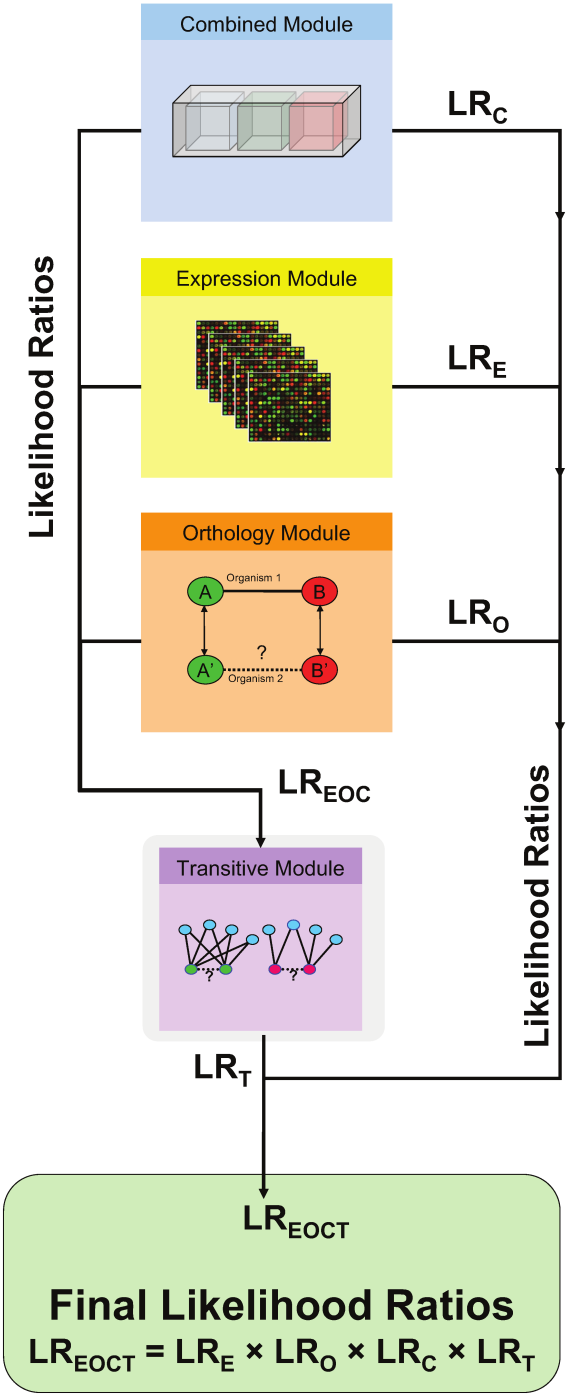
The PIPs framework was developed within the Barton Group (Scott and Barton, 2007) to identify and rank predicted human protein-protein interactions. The framework used a semi naïve Bayesian method to combine multiple sources of evidence to predict a likelihood of interaction and predicted 37,606 protein-protein interactions. Evidence included correlation of gene expression; orthology; annotations, including co-occurrence of domains, post translational modifications and co-localisation; and a network module that considered the local network of interactions. The sources of evidence were grouped into 4 modules that calculate a likelihood ratio of interaction for each protein pair. Figure 1.8 shows the four modules and how their likelihood ratios were used to calculate the final likelihood ratio for each protein pair.

1.8.1 Module Structure and Calculating Likelihood Ratios

Modules were built to assign a likelihood ratio of interaction for given protein pairs. This was done by assigning a score to a protein pair based on a set of evidence. The score was discretised to one of a number of bins. Training the modules involved computing a likelihood ratio for each bin given the ratio of positive to negative protein pairs assigned to that bin. During testing, protein pairs received the likelihood ratio of the bin to which they were assigned based on their evidence.

For example the gene expression module had 20 bins, assigned values ranging from -1 to 1. During training, for each training set example, both positive and negative, a Pearsons correlation of co-expression was calculated based on experimental evidence. The training example was then assigned to the bin corresponding to their

Figure 1.8: PIPs Framework Version 1. Each module is indicated by a coloured box (Blue, yellow, orange or purple). The arrows indicate how the likelihood ratios calculated by each module are combined. The final likelihood ratio for each protein pair is the product of the likelihood ratios calculated by each module. The Transitive module uses the product of the likelihood ratios from the Combined, Expression and Orthology modules for each protein pair to generate the local network of interactions.



Pearson correlation value and counted in this bin. The ratio of the proportion of positives to the proportion of negatives in a given bin was used to calculate the likelihood ratio for that bin.

Modules

Expression Module The Expression module used dataset GDS596 (Su et al., 2004; Barrett et al., 2007) and the Pearsons correlation to calculate a level of co-expression between protein pairs. The Pearsons correlation was discretised into 20 bins.

Orthology Module The Orthology module worked on the hypothesis that a protein pair is more likely to interact if orthologous proteins in another species are known to interact. The Orthology module used InParanoid (Berglund et al., 2008) to map from proteins in the human proteome to proteins in Yeast, Worm and Fly. Protein-protein interactions for pairs of proteins in other species were derived from BIND (Alfarano et al., 2005), DIP (Salwinski et al., 2004) and GRID (Breitkreutz et al., 2003) (now BioGRID). The Orthology module calculated a likelihood ratio based on the InParanoid scores for each of the proteins and their orthologues and whether an interaction in an orthologous species is known to exist.

Combined Module The Combined module incorporated 3 types of annotation:

1. Localisation
2. Domain co-occurrence
3. Post translational modifications

These were combined in a 3 dimensional matrix. For co-localisation protein pairs were assigned to one of 4 bins dependent on if they were located in the same, neighbouring, different or unallocated compartments. Domain co-occurrence was based on InterPro (Mulder et al., 2007) and Pfam (Finn et al., 2006) domains. A Chi square score was calculated for the co-occurrence of two domains being part of an interacting pair of proteins. The score assigned to Co-occurrence of post translational modifications was calculated based on Equation 1.8.1.

$$PTM_{score} = \frac{P(PTM[i], PTM[j]|I)}{P(PTM[i]|I) \times P(PTM[j]|I)}$$

Equation 1.8.1: PTM Scoring function

where $PTM[i]$ and $PTM[j]$ are unique post translational modifications and I is the set of interacting proteins available during training. This score was discretised to calculate and assign likelihood ratios.

Transitive Module The transitive module was based on the hypothesis that two proteins are more likely to interact if they share an increasing number of common protein interactors. The score was based on the local topology of the predicted interaction networks calculated by the Expression, Orthology and Combined modules. The product of each of the modules for each protein pair was used to construct the predicted interaction network and edges were filtered for likelihood ratios above 10. A topology score was calculated based on Equation 1.8.2.

In Equation 1.8.2 E_c is the set of edges that connect proteins i and j to common interactors, E_i is the set of edges that involve protein i and s_e is the likelihood ratio of edge e and $E_i \setminus E_c$ is the set difference between E_i and E_c. The module did not

$$T_{score} = \frac{\sum_{e \in E_c} s_e}{1 + |E_i \setminus E_c| + |E_j \setminus E_c|}$$

Equation 1.8.2: PTM Scoring function

consider the score between the two proteins being investigated, only the common neighbours of the two proteins.

1.9 Scope of This Thesis

The purpose of this thesis is to investigate the prediction of protein-protein interactions in humans with the PIPs predictor. Chapter 2 describes the development of two new modules (Clustering and Sequence Analysis modules, Sections 2.2.2 and 2.2.4 respectively) and improvements to the PIPs 1 predictor by the introduction of the Gene Ontology to the Combined module and analysis of different gene expression datasets (Section 2.2.1 and 2.2.3). Chapter 3 brings together the modules developed in Chapter 2 to create the PIPs 2 predictor and then analyses its accuracy. Chapter 4 looks at the predictions that were made by the PIPs 2 predictor. Chapter 5 talks about the web services that have been developed during the thesis. Section 5.2 is about the PIPs webserver (McDowall et al., 2009), which makes the predictions that were made by the PIPs 1 predictor publicly available to search and download. Section 5.3 describes the development of an interface to the PIPs predictions by the FuncNet project.

Chapter 6 discusses the application of the PIPs 2 predictor in other organisms and goes on to investigate the potential for training the PIPs 2 predictor in one organism and applying the trained model to a second organism that would not

have enough training data available. Chapter 7 accesses the accuracy of the Jpred predictor, for which the predictions were used by the Sequence Module (see Section 2.2.4) and goes on to determine whether it is possible to predict the accuracy of a secondary structural prediction. Chapter 8 highlights the major conclusions of the thesis and describes the improvements and paths that the project could take in the future

Chapter 2

Module Development

Preface

This chapter covers the development and enhancement of modules that are considered by the PIPs 2 predictor. Section 2.2 describes the new modules that make predictions based on clustering of predicted protein interaction networks (Section 2.2.2) and sequence analysis (Section 2.2.4) as well as improved modules from PIPs 1 that were significantly changed, such as the Expression module (Section 2.2.3) and Combined module (Section 2.2.1).

2.1 Introduction

This chapter describes the methods that have been applied to improve the predictions of the PIPs predictor. Techniques include the use of semantic similarity of Gene Ontology terms, clustering of the predicted interactome, analysis of the use of gene expression and sequence analysis. the individual modules are each introduced and described separately in this Chapter.

2.1.1 Combined Module and The Gene Ontology

The initial PIPs 1 Combined module calculated likelihood ratios based on the co-occurrence of InterPro domains and post translational modifications and the co-localisation of proteins pairs. To identify potential improvements the use of the Gene Ontology was also investigated for inclusion in the Combined module.

The Gene Ontology (GO) is a hierarchical vocabulary of terms that are used to describe the roles of genes and gene products (Ashburner et al., 2000). Proteins may be assigned one or a number of terms from each of the three branches of the GO to describe Molecular Function (F), Cellular Compartment (C) or Biological Process (B).

The GO is structured as a Directed Acyclic Graph (DAG). With a measure of semantic similarity it is hypothesised that protein pairs that have assigned terms that are closer in the DAG are more likely to interact than those annotated with unrelated terms. Many solutions have been provided for measuring semantic similarity (Jiang and Conrath, 1997; Lin, 1998; Resnik, 1995). Each measure considers the nearest common term, semantic distance between terms and the information content (IC) of the nearest common term where IC is a measure of how general a term is. Lord et al. (2003) applied each of the semantic measures for use with the GO to measure protein sequence similarity in relation to each of the three branches of the GO. It was found that the sequence similarity of a protein was linked with the assigned Molecular Function term, but sequence similarity and semantic similarity for Biological Process and Cellular Compartment were not related.

To account for multiple common disjunctive ancestors Couto et al. (2007) proposed a modification to each of the three semantic measures (Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998) known as GraSM. It was found that the most effective measure of similarity between two proteins was to apply GraSM with the Jiang and Conrath (1997) measure of similarity. The results are described in Section 2.2 and the results in Section 2.3.1.

2.1.2 Cluster Module

Protein-protein interaction networks can be viewed as graphs, where proteins are represented by nodes and edges between nodes represent the presence of an interaction between the corresponding pair of proteins. It is then possible to apply graph theory approaches, such as node degree, clustering coefficients and topology analysis to characterise the graph. Such analysis has been applied to numerous interaction networks, most of which have a power law degree distribution and a high clustering coefficient (Ravasz et al., 2002; Barabasi and Oltvai, 2004; Gandhi et al., 2006; Jeong et al., 2000; Wagner and Fell, 2001). A hub, or modular, centric view has been found to be an ideal way to view biological networks (Vallabhajosyula et al., 2009), these hubs/modules are representative of biological complexes of functional pathways. This concept of being able to identify clusters of proteins within a network based on the topology has been used by many groups for the prediction of complexes (Liu et al., 2009; Bader and Hogue, 2003; King et al., 2004) and functional pathways. It is also possible to use the structure of the network to predict missing links (Clauset et al., 2008).

Cluster algorithms identify groups of entities within a set such that entities that

are more similar are associated together and therefore apart from entities that are less similar. Algorithms can then group entities by calculating a distance metric based on each entities properties and from this subset the entities with similar properties. Metrics of distance can include Euclidean, Manhattan or Hamming distances. However, clustering a graph is dependent on the weight of the edges and the topology of the network. When it comes to graphs, the edges can be assigned weights dependent on given properties, but they can also be a binary representation of whether there is an interaction or not. Once a distance has been calculated it is then possible to cluster the entities. There are numerous clustering methods that have been developed and applied to identify complexes and functional modules within interactomes, such methods include NeMo (Rivera et al., 2010), MCODE (Bader and Hogue, 2003), MCL (Dongen, 2000; Enright et al., 2002), RNSC (King et al., 2004) and SPC (Blatt et al., 1996).

Section 2.2.2 describes the investigation of the hypothesis that pairs of proteins that are clustered together are more likely to interact than pairs of proteins that are located within different clusters.

2.1.3 Expression Module

Predicting protein-protein interactions based on gene expression data is not a new concept and has been found to have varying degrees of success (Jansen et al., 2002; Bhardwaj and Lu, 2005; Rhodes et al., 2005; Ramani et al., 2008). Often it is found that co-expressed genes express stable interacting protein pairs and are part of complexes, such as the proteasome (Jansen et al., 2002).

Section 2.2.3 describes the improvement of the Expression module within the

PIPs framework and the use of gene expression data for the prediction of protein-protein interaction prediction. The use of different measures of correlation, the effect it has on the prediction of protein-protein interaction and the incorporation of different gene expression datasets were investigated.

2.1.4 Sequence Module

The modules that have been described previously for the PIPs 1 framework (see Section 1.8) all rely on experimental evidence or the presence of annotations. Even the Transitive and Clustering modules can only make predictions if there is strong enough evidence for the presence of an interaction and their information depends on the Expression, Orthology and Combined modules. The ability to use protein primary sequence data would mean that the only limitation would be to have a fully sequenced genome rather than the limitations with technologies and number of annotations. Protein domains are already considered as part of the Combined module and have been shown previously to be predictive in classifying protein-protein interactions (Sprinzak and Margalit, 2001). However, although the use of domains has been found to be predictive, it is dependent on the presence of annotations to identify the domains. Therefore sequence representation (Martin et al., 2005; Shen et al., 2007) and amino acid composition (Roy et al., 2009) have been developed to make predictions that are independent of domains and annotations.

Section 2.2.4 investigates the implementation of sequence based methods that are not reliant on the presence of annotations of domains. This method is therefore limited only by number of proteins that have been sequenced rather than the annotations present within a database.

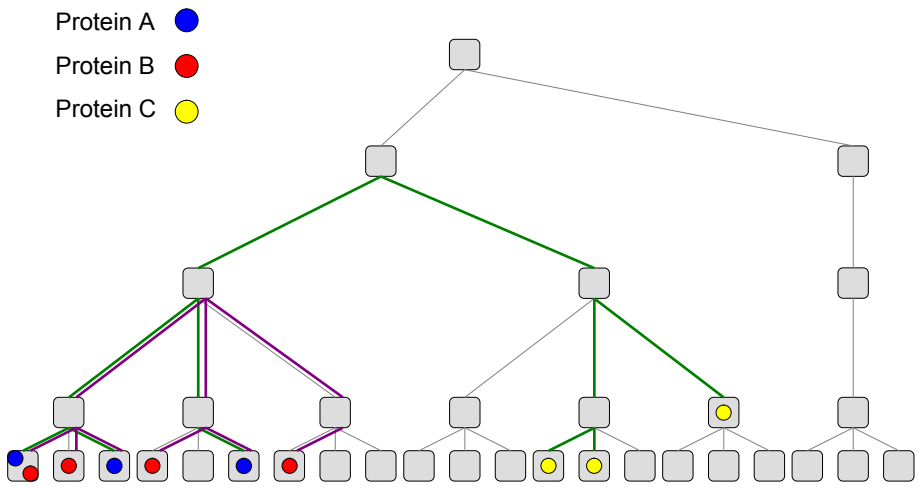
2.2 Methods

2.2.1 Analysis of Annotated Gene Ontology Terms as Part of the Combined Module

Calculation of Semantic Similarity

For the prediction of protein-protein interactions, the Jiang Conrath semantic similarity (Jiang and Conrath, 1997), with the GraSM adjustment (Couto et al., 2007), was selected for measuring the semantic similarity between GO terms annotating two proteins. To measure the similarity between two proteins the frequency of each GO term is calculated, this represents the probability of a randomly selected protein having a given GO term, see Figure 2.1. The frequency of terms is dependent on the number of proteins within the proteome associated to each term, as described in Equation 2.2.1 where t is a term within the GO, $Count(t)$ is the number of proteins assigned term t as the most informative term and C_t is the set of all children of t.

Figure 2.1: The figure depicts a hierarchical directed acyclic graph of terms (grey boxes), where each term below the root term (top grey box) is more specific. Proteins A (●), B (●) and C (●) are each assigned N terms (A: 3; B: 4; and C: 3). The semantic distance between Protein A and B is highlighted in purple and the distance between protein A and C is highlighted in green, therefore showing that terms assigned to Protein A and B are closer than between A and C.



$$Freq(t) = Count(t) + \sum_{i \in C_t} Count(t_i)$$

Equation 2.2.1:

The probability of randomly selecting term t is estimated using Equation 2.2.2, where the *maxFreq* is the frequency of the root term for each GO branch.

$$Prob(t) = \frac{Freq(t)}{maxFreq}$$

Equation 2.2.2:

The Information Content IC value of term t is the negative log of its probability, Equation 2.2.3 (Couto et al., 2007).

$$IC(t) = -\log(Prob(t))$$

Equation 2.2.3: Information Content

$Share$ is the IC value of the common ancestor between two terms, t_1 and t_2. If two terms have multiple common disjunctive terms, the average for all the IC values for each common disjunctive ancestor is taken, where a represents all the possible common disjunctive ancestors of the terms t_1 and t_2, Equation 2.2.4 (Couto et al., 2007).

The similarity between two terms based on the Jiang calculation is shown in Equation 2.2.5 (Couto et al., 2007).

Within the PIPs framework, a semantic similarity module was created to calculate the similarity between the GO terms of two proteins by considering Molecular Function, Cellular Compartment and Biological Process separately. This allowed

$$Share_{GraSM}(t_1, t_2) = \overline{IC(a) \mid a \epsilon CommonDisjAnc(t_1, t_2)}$$

Equation 2.2.4: Information Content of the nearest common ancestor

$$Sim_{JCGraSM}(t_1, t_2) = \frac{1}{IC(t_1) + IC(t_2) - 2 \times Share_{GraSM}(t_1, t_2)}$$

Equation 2.2.5: Similarity between two terms

the analysis of each of the branches independently. The calculated semantic similarities were grouped into 3 bins; < 0.2, ≥ 0.2 to < 0.6 and ≥ 0.6. The analysis was performed for positive and negative datasets from which a likelihood ratio (LR) can be calculated for each of the groups using Equation 2.2.6, as described in Chapter 1.6.

$$LR = \frac{P(f \mid I)}{P(f \mid \sim I)}$$

Equation 2.2.6: Prior probability ratio

where the likelihood ratio for a feature (f) allocated to a bin, is the ratio of observed positives divided by the number of observed negatives given a training set of known positives (I) and negatives ($\sim I$).

The three GO branches were considered independently and in combinations using full Bayesian analysis with each of the features used in the Combined module (Post-translational modification, Co-occurrence of domain and co-localisation) (Scott and Barton, 2007). Bayesian Information Criterion (BIC) scores (Schwarz, 1978) were calculated to determine the most viable of the combinations. The BIC score is calculated using Equation 2.2.7.

where n is the number of observations, k is the number of groupings (bins) and

$$BIC = -2 \times \ln L + k \ln(n)$$

Equation 2.2.7: Bayesian Information Criterion

L is the maximum Likelihood ratio value for the estimated model.

Co-occurrence of Protein Domains and Post Translational Modifications

The Domains sub module calculates a likelihood of interaction based on the co-occurrence of InterPro and Pfam domains between a pair of proteins by calculating a Chi square score. The same method was also used for the co-occurrence of post-translational modifications between a pair of proteins. Further details are provided as part of Chapter 1.8.1.

Data Source

The GO term annotations were downloaded from the Gene Ontology Annotation website (October 2007) and loaded into the database. All annotations for GO terms were loaded for all three trees within the GO (Biological Process, Cellular Compartment and Molecular Function). Three types of relationships are defined within the GO to annotate the association between terms, *is_a*, *part_of* and *regulates*. Calculations described here use only the *is_a* reference as in the case of:

$$A \xrightarrow{is\_a} B$$

defines that term A is a subset of GO term B within the GO hierarchy. Because the *is_a* relationship is transitive it means that if A *is_a* B and B *is_a* C then A *is_a* C. *Part_of* in the case of:

$$A \overset{part\_of}{\longrightarrow} B$$

defines that GO term A can be a subset of B, but not all of B will be a part of A.

2.2.2 Clustering of Protein Interaction Networks

Clustering the Protein-Protein Interaction Network

This section investigates the use of the MCL algorithm to cluster the predicted protein-protein interaction network with the aim of identifying clusters of proteins that are likely to interact. The integration within the PIPs framework is also investigated. Edge weights are represented by the product of the likelihood ratios calculated by the Expression, Orthology and Combined modules within the PIPs framework (LR_{EOC}).

MCL (Dongen, 2000) was chosen over that of hierarchical clustering and k-means methods for several reasons. Both hierarchical and k-means clustering require prior information about the network, or optimisation to determine representative values. MCL identifies clusters based on the intrinsic properties of the network, rather than prior knowledge about the number of clusters or selection of tree threshold points. MCL has also been shown to be effective in the clustering of protein-protein interaction networks (Brohee and van Helden, 2006).

Clustering is performed on the protein pairs within the training set. All positive and negative training examples are included in the clustering as long as they have an LR_{EOC} value greater than a set threshold.

Accuracy Measurement

To analyse the similarity between the clusters generated by MCL and known complexes the matching statistics, Accuracy and Separation, as defined by Brohee and van Helden are used (Brohee and van Helden, 2006). The calculations are based on an n by m contingency table (T) for n complexes and m clusters, where the i^{th} row relates to a specific complex and the j^{th} column relates to a given cluster and each cell within the table is the number of proteins that are common between the i^{th} complex and and the j^{th} cluster.

Positive Predictive Value The positive predictive value (PPV) is the proportion of proteins that are part of complex i and are present in cluster j relative the size of cluster j (Equation 2.2.8) (Brohee and van Helden, 2006).

$$PPV_{i,j} = \frac{T_{i,j}}{\sum_{i=1}^{n} T_{i,j}} = \frac{T_{i,j}}{T_{.j}}$$

Equation 2.2.8: Positive predictive value (PPV)

where T_j is the number of proteins in cluster j and $T_{i,j}$ is the number of proteins in cluster j and part of complex i. As some proteins can belong to multiple complexes, a cluster-wise PPV value (PPV_{cl_j}) is the maximum proportion of proteins in a set of clusters. The general PPV is calculated with Equation 2.2.9 (Brohee and van Helden, 2006).

Sensitivity Sensitivity is the fraction of proteins in complex i that are present in cluster j (Equation 2.2.10) (Brohee and van Helden, 2006) where N is the number of proteins that belong to complex i.

$$PPV = \frac{\sum_{j=1}^{m} T_{.j} PPV_{cl_j}}{\sum_{j=1}^{m} T_{.j}}$$

Equation 2.2.9: general equation for the Positive Predictive value

$$Sn_{i,j} = \frac{T_{i,j}}{N_i}$$

Equation 2.2.10: Sensitivity

The complex wise sensitivity (Sn_{co_i}) is the maximum sensitivity for a complex for a given set of clusters. The clusterwise sensitivity is then defined with Equation 2.2.11 (Brohee and van Helden, 2006).

$$Sn = \frac{\sum_{i=1}^{n} N_i Sn_{co_i}}{\sum_{i=1}^{n} N_i}$$

Equation 2.2.11: Cluster-wise sensitivity

Accuracy Accuracy is a measure of how closely the clusters resemble known complexes. As the number of proteins within a cluster that match a complex increases, the accuracy increases. The perfect score (1) is achieved when all proteins from a complex are in the same cluster.

Accuracy is the geometric mean of the cluster-wise sensitivity and the positive predictive value (Equation 2.2.12) (Brohee and van Helden, 2006).

Geometric Separation The geometric separation is a measure of the proportion of elements in cluster j found in complex i by the proportion of elements in complex

$$Acc = \sqrt{Sn \cdot PPV}$$

Equation 2.2.12: Accuracy

i found in cluster j. The perfect score, 1.0, is obtained if the cluster contains only the proteins of a single complex, however the complex may be split over several clusters(Brohee and van Helden, 2006).

While the $PPV_{i,j}$ (Equation 2.2.8) is a cluster-wise measure, the complex-wise measure is shown in Equation 2.2.13 (Brohee and van Helden, 2006).

$$F_{row_{i,j}} = \frac{T_{i,j}}{\sum_{j=1}^{m} T_{i,j}}$$

Equation 2.2.13: Complex-wise Positive Predictive Value

The separation, $Sep_{i,j}$, is calculated by Equation 2.2.14 (Brohee and van Helden, 2006).

$$Sep_{i,j} = PPV_{i,j} \cdot F_{row_{i,j}}$$

Equation 2.2.14: Separation

Complex and cluster-wise separations are therefore calculated with equation 2.2.15 and Equation 2.2.16 respectively (Brohee and van Helden, 2006).

$$Sep_{co_i} = \sum_{j=1}^{m} Sep_{i,j}$$

Equation 2.2.15: Complex-wise Separation

$$Sep_{cl_j} = \sum_{i=1}^{n} Sep_{i,j}$$

Equation 2.2.16: Cluster-wise Separation

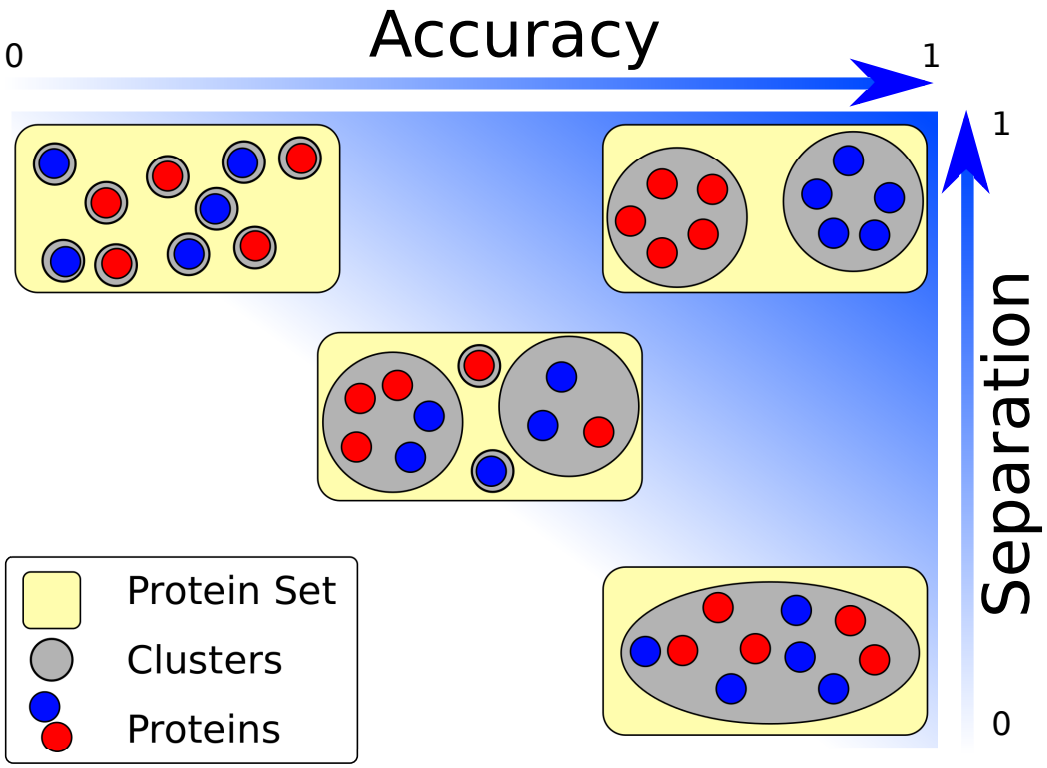
The final geometric separation is calculated with Equation 2.2.17 (Brohee and van Helden, 2006).

$$Sep_{cl_j} = \sqrt{\frac{\sum_{i=1}^{n} Sep_{co_i}}{n} \cdot \frac{\sum_{j=1}^{m} Sep_{cl_j}}{m}}$$

Equation 2.2.17: Geometric Separation

Accuracy and Separation The schematic in Figure 2.2 shows the aim of a clustering algorithm to optimise for both the accuracy and separation scores. Both

Figure 2.2: The matching statistics Accuracy and Separation (Brohee and van Helden, 2006). The aim is for the maximisation of both statistics to obtain a perfect separation of the proteins such that the clusters match the biological complexes.



measures (Accuracy and Separation) have their own disadvantage when considered in isolation:

- The accuracy measure will provide a perfect score if all proteins are grouped into the same cluster.
- The separation measure will provide a perfect score if all proteins are grouped individually such that all clusters contain only a single protein.

Due to the disadvantages of both measures, the real benefit is when they are analysed together.

Calculating the Cluster Score

After clustering based on the product of the likelihood ratios from the Expression, Orthology and Combined modules (LR_{EOC}), each cluster is then scored. For each cluster (C_x), all possible interactions I_x, are allocated a value of 1, unless they have a protein pair present within the training set, in which case they are allocated the LR_{EOC} value, c. The number of possible interactions, N_x, is calculated via Equation 2.2.18. The Cluster Score for each cluster is calculated by Equation 2.2.19.

$$N_x = \frac{n(n-1)}{2} \qquad \text{where} \qquad n = |C_x|$$

Equation 2.2.18: Calculate the number of edges in a complete clique

$$C_{score} = \frac{\sum_{i \in I_t} S_i}{N_x}$$

Equation 2.2.19: Cluster Score

where $S_i = LR_{EOC}$ if i is an element of I_t where I_t represents all protein pairs in the positive and negative training sets. Otherwise $S_i = 1$.

For clusters that contain a large number of proteins with only a few strong interactions, the Cluster score, C_{score}, will be lower than a smaller cluster with the same number of strong interactions.

Calculating the Cluster Likelihood Ratio

The likelihood ratio for a protein pair is assigned based on whether the two proteins of the pair are within the same cluster. If a protein pair is present within the same cluster, they are assigned to one of B bins dependent on the Cluster Score. If the

protein pair is not within the same cluster they are assigned to a separate bin.

The likelihood ratio for each bin is calculated based on the allocation of positive and negative examples from the training set. The likelihood ratio (LR_M) is calculated for each bin based on Equation 2.2.20.

$$LR_M = \frac{P_i \div P}{N_i \div N}$$

Equation 2.2.20: Likelihood ratio

Where P is the number of positive training examples and N is the number of negative training examples. P_i and N_i are the number of positive and negative examples, respectively to be allocated to bin i.

Independence of Predictions

As stated in Chapter 1.6, it is important that the predictions that are made by each module are independent within a naïve Bayes network. Testing for independence between the predictions made by the Clustering and Transitive modules is important in determining whether both modules can be included in the final predictor as they both use the same initial predicted network of protein-protein interactions for calculating the likelihood ratio of interaction for each protein pair. Independence between the Clustering and the Transitive module can be tested using Pearson's correlation coefficient. If both of the modules are not independent then there is a chance of artificially enhancing the likelihood of two proteins interacting. In this case it would be possible to include both methods within the final predictor as they are making independent predictions on the data. If there is a high correlation between the two predictive modules then two final likelihood ratios would be calculated

(LR_{EOCT} and LR_{EOCM}).

2.2.3 Analysing Gene Co-expression

Gene Expression Data

Gene expression data can be downloaded from numerous repositories, most notably ArrayExpress (Parkinson et al., 2009) at the EBI and GEO (Barrett et al., 2007) at the NCBI. For this analysis the expression datasets were downloaded from the ArrayExpress repository. The following gene expression datasets were downloaded:

- E-TABM-145 (Su et al., 2004) - This dataset was used by the PIPs version 1 framework.
- E-GEOD-7307 (Release by Roth 2007)
- E-GEOD-3526 (Roth et al., 2006)

E-TABM-145 uses the A-AFFY-33 chip, which has probes for 13,639 distinct proteins. In comparison, the E-GEOD-7307 and E-GEOD-3526 datasets use the A-AFFY-44 chip that has probes for 18,334 distinct proteins.

The probe expression values were downloaded from ArrayExpress as CEL files which are the original output files from the microarray machine. The values in the CEL files have not been subjected to statistical normalisation. By downloading the original files, multiple datasets can be joined prior to normalisation. This ensures that all the data is treated identically rather than downloading the pre-filtered expression values which could have been treated differently before being submitted to the ArrayExpress database.

In addition to the above datasets, a subset of 500 probes from the gene expression datasets of E-GEOD-4295, E-GEOD-8799, E-GEOD-9217, E-GEOD-11651 and E-GEOD-12222 all based on the AffyMetrix gene chip A-AFFY-47 were combined and normalised using Robust Multichip Average (RMA) (see Section 2.2.3 - Data Normalisation) to investigate the effect of different measures of correlation. The reason for using a subset of the probes was to make the calculations more tractable within a reasonable time frame.

Data Normalisation

All the gene expression datasets considered were generated using the AffyMetrix GeneChips which consist of two types of probes. The two types of probes are:

1. Perfect Match (PM): These probes are perfectly complementary to their corresponding genes.

2. Mismatch (MM): These probes have a single base pair change in the middle of the sequence.

Typically there are between 16-20 probe pairs per gene, although for the HG-U133 (A-AFFY-44) array there are only 11 pairs. It is the fluorescent values from these probes that are used to calculate the level of expression of a particular gene. However before an expression value can be calculated, the data for all experiments and repeats needs to be normalised to account for variances in the running of the experiment. The most common procedures are:

1. Average Difference (AvgDiff)

2. Model Based Expression Difference (MBED) (Li and Hung Wong, 2001; Li and Wong, 2001)

3. MAS 5.0 Statistical Algorithm (Affymetrix, 2002)

4. Robust Multichip Average (RMA) (Irizarry et al., 2003)

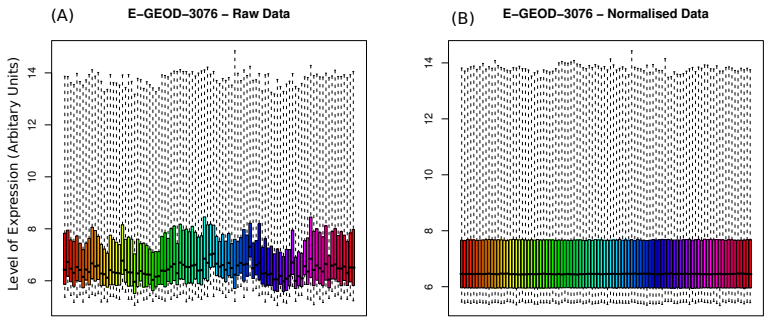
The results of a later study by (Bolstad et al., 2003) demonstrated that the most informative method for the normalisation of multichip microarray sets was RMA (Irizarry et al., 2003). RMA approach is to model the distribution of the perfect match probes and then uses quantile normalisation.

The RMA method has been incorporated within the affy package within the BioConductor project (Gentleman et al., 2004). All of these tools are available within the R programming language (R Development Core Team, 2009).

The expression datasets described above were thus normalised using RMA within R. These datasets were then formatted and inserted into a MySQL database ensuring a correct mapping between the arrays, probes and transcripts.

The effect of data normalisation is shown in Figure 2.3. Figure 2.3A illustrates the pre-normalisation values for each of the experiments within the yeast dataset E-GEOD-3076. As shown in Figure 2.3B, after normalisation, by accounting for variances in the experiments, all of the expression sets fall within the same data range. As a consequence, no experiment over shadows another due to variances that occurred during the experiment. The only variance is the difference in gene expression.

Figure 2.3: Normalisation of Yeast gene expression dataset E-GEOD-3076. A: Before normalisation; B: After normalisation.



Calculation of Gene Co-expression

The purpose of this study was to identify the differences in the correlation of gene expression dependent on the method applied and to identify the best method for the calculation of correlation for the prediction of protein-protein interaction.

Four measures of correlation were studied:

1. Pearson Rank Correlation Coefficient (Pearson, 1895) and later reviewed (Rodgers and Nicewander, 1988)

2. Spearmans Rank Correlation Coefficient (Spearman, 1904)

3. Kendall Rank Correlation (Kendall, 1938)

4. Hardin (Hardin et al., 2007)

Pearsons Rank Correlation Coefficient is a holistic method that considers all points within the datasets. However Spearmans, Kendall and Hardin are more robust methods of measuring correlation as they are not affected by extreme outliers within the data. Pearsons can be easily polluted by spurious points that are significantly different from the other data points; as a result this can lead to correlations

$$r = \frac{\sum_{i \in N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i \in N}(X_i - \overline{X})^2 \sum_{i \in N}(Y_i - \overline{Y})^2}}$$

Equation 2.2.21: Pearsons Rank Correlation Coefficient

being identified where in actual fact there are none. The Pearson Rank Correlation Coefficient is shown in Equation 2.2.21 where N is the set of X, Y values and i is a reference to a specific pair of X and Y within N. From the equation it is possible to identify why this measure of correlation is affected by extreme outliers. The use of the mean of X and Y will be heavily biased if there are extreme values present within the data.

Kendall and Spearman's both rank the data and are a measure of the differences between the ranking of the X and Y values. The equations for Kendall and the Spearman Rank Correlation Coefficient are shown in Equation 2.2.22 and Equation 2.2.23 respectively.

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

Equation 2.2.22: Kendalls Tau

$$\rho = 1 - 6\sum \frac{d^2}{n(n^2-1)}$$

Equation 2.2.23: Spearmans Rank Correlation Coefficient

Where n in the number of data points, n_c is the number of concordant data points, n_d is the number of discordant data points and d is the difference between X_i and Y_i. In Equation 2.2.22, when the sets of X and Y are ranked, if X_i and Y_i are ranked at the same point, this is known as a concordant point, discordant points are where two points are ranked differently.

Both Kendall and Spearman's use a ranked measure of correlation making them less susceptible to outliers within the data. But it is this benefit that can also be a deficit especially as it means that not all points within the data are considered. Within gene expression this means that an interesting result might become overlooked. However in protein-protein interaction prediction it is the trend over the whole series that is important.

Hardin et al. (2007) use a different method to calculate correlation. Hardin et al. (2007) implements an M-estimator to calculate the scatter of a 2 dimensional plot then use this to calculate a biweight correlation (Hardin et al., 2007).

Database

All of the gene expression data, along with the pre-calculated correlation values of all gene combinations and all probe metadata are held on a MySQL 5.0.43 database. The total storage space required for each of the gene expression datasets and supporting data is 75Gb.

2.2.4 Protein Sequence Analysis

Protein Feature Representation for Support Vector Machines (SVMs)

Support vector machines (SVMs) are ideal for the classification of data that can be represented by multiple feature vectors. The original algorithm designed by Vapnik in 1963 was a linear classifier. It wasn't until later that the kernel trick was applied to create a non-linear classifier (Boser et al., 1992), where the kernel trick is the mapping of real observations into a multidimensional space such that the representation of the data has a linear structure.

SVMs were employed to predict the likelihood of interaction based on a set of features that will represent each protein. The SVM classifier then provided a classification of 1 or -1 depending on whether the proteins are predicted to interact or not. An extra feature that measures the correlation of the features between the two proteins (Pearson (Rodgers and Nicewander, 1988) or Spearmans (Spearman, 1904)) was also investigated.

In total three SVMs were investigated, each representing different feature sets. They include:

Tripeptide Secondary Structure Motif The first feature to be investigated was elements of secondary structure as predicted by Jpred, a secondary structure predictor (Cole et al., 2008). Tripeptide motifs of the secondary structure were used where each amino acid was represented by either "E" (β-sheet), "H" (α-helix) or "-" (unspecified), for each protein (eg EEE, EEH, EHE). A tripeptide motif was used rather than a larger motif as the number of potential motifs grows exponentially resulting in incomplete coverage and therefore overfitting of the training data. With a tripeptide motif there are 27 potential features to describe each protein. A reduced motif set that has only 10 features to represent a protein was also investigated. While EEH, EHE and HEE are equivalent in the reduced set, they are separate features within the full set. The reduced set is therefore position insensitive and more applicable to the proportions of the secondary structure features than the position within the sequence. Each feature was encoded as the sum of the occurrence of each motif within the protein.

Tripeptide Sequence Motifs The method developed by Shen et al. (2007) was implemented with a reduced amino acid alphabet to represent each of the residues within a protein. They reduced the number of amino acids from 20 down to 7 different residue types. This means that rather than having 8000 tripeptide motifs to represent a single protein, there are 343 motifs with the reduced alphabet. The reduced alphabet is shown in Table 2.1.

As with the reduced Jpred motif, a reduced sequence feature vectors for the reduced alphabet was investigated. For example $\alpha\beta\gamma$, $\alpha\gamma\beta$, $\gamma\alpha\beta$, $\gamma\beta\alpha$, $\beta\gamma\alpha$ and $\beta\gamma\alpha$ are equivalent within the reduced tripeptide feature set, whereas in the full set they are distinct features. Therefore in the reduced feature set there are 84 motifs rather than 343.

Proportions This feature set represents each protein with the proportion of 14 different elements. These include:

- Predicted alpha helix (Cole et al., 2008)
- Predicted beta sheet (Cole et al., 2008)
- Predicted not alpha or beta regions (Cole et al., 2008)

Table 2.1: Amino acid sub-classification adapted from (Shen et al., 2007).

New Residue	Represented Amino Acids
α	A, G, V
β	F, I, J, L, P
γ	M, S, T, Y
δ	H, N, Q, W
ϵ	K, R
ζ	D, E
η	C

- Predicted solvent accessibility (0-5%, 5-25%, >25%) (Cole et al., 2008)
- One feature for each of the 7 amino acids of the reduced primary structure for the protein
- Predicted proportion of the protein that is disordered based on the predictions made by the VSL2B predictor (Peng et al., 2006).

Prediction of Protein Secondary Structure

Secondary structur predictions were made for all proteins within the PIPs v1 database so that it would be technically possible to make a full set of binary protein pair predictions. There were 66,654 sequences proteins in the PIPs (v1) database were downloaded from the IPI database in February 2007; Jpred 3 was implemented to predict protein secondary structures (Cole et al., 2008). The sequences were filtered for redundancy using BLAST (2.2.13) to align the proteins within the PIPs database. The redundancy threshold was set at 95% sequence identity and 95% sequence length as reported by BLAST. This left 49,892 sequences that were passed to Jpred to predict secondary structure. Of the 49,892 sequences 3488 proteins that had known structures within the PDB, these were also submitted to Jpred. The reason for submitting these known structures to Jpred is so that all sequences were treated identically.

This resulted in 48,122 Jpred predictions that mapped to a further 12,932 proteins based on sequence redundancy giving a total of 61,054 proteins with an associated predicted secondary structure. 1770 of the sequences failed to generate a prediction and so these have been left out. This meant that there are 5600 proteins

without a predicted secondary structure. The reason for sequences that lack a predicted secondary structure is due to lack of PSI-BLAST results during the Jpred predictions of the secondary structure.

The output of Jpred predictions for each amino acid within the protein primary structure includes the secondary structural feature "E" (β-Sheet), "H" (α-helix) or "-" (neither β-sheet or α-helix). Jpred also provides three levels for the predicted burial of the residue within the tertiary structure.

SVMs

The kernlab (Karatzoglou et al., 2004) package available within the R programming language (R Development Core Team, 2009) contains methods for training and classifying data using Support Vector Machines. The package was developed in R so it is capable of incorporating custom kernels within the SVMs. The kernels that are already implemented within kernlab package that were investigated include:

- Gaussian Radial Base Function (RBF)
- Polynomial
- Vanilla (Linear kernel)
- Hyperbolic tangent
- Laplace
- Bessel
- ANOVA RBF

The SVMs were trained with 1000 positive and 1000 negative randomly selected examples of interacting protein pairs. To test the SVMs, a blind set of 33094 positive and 33094 negative examples was used. A variety of cost (C) thresholds were also set, these affect the final classifications. The cost function measures were set at 20, 15, 10, 5, 1 0.5 and 0.1.

To validate the predictions, the SVMs were also trained with 2000 random protein pairs. Random values were assigned for each of the feature sets with ranges between 0 and 1. The pseudorandom number generator used is the Mersenne Twister, which has a period of $2^{19937} - 1$ (Matsumoto and Nishimura, 1998). The SVMs are then trained with the random examples and the results compared to SVMs trained with known examples. The purpose is to identify whether the SVMs trained with real examples are actually more effective than those trained on random values.

Data Normalisation

SVMs do not handle raw data well, they are more effective once the data has been normalised. Two methods of normalisation were investigated:

Scaling All of the values for each feature were scaled across the full set of the proteome. Where 1 was allocated to the largest value within the dataset and 0 to the lowest score, the rest of the scores were then linearly scaled as appropriate.

Standardisation This replaces the original value with the z-score:

$$x'_n = \frac{x_n - \mu}{\sigma}$$

Equation 2.2.24:

where μ is the mean and σ it the standard deviation of the full set of features of a vector for each protein, x_n is the element for the individual protein n and x'_n is the normalised value for protein n.

Quality Assessment

The following are calculations that can be derived from the results of the test sets based on the classification of the positive (P) and negative (N) examples as either true positive (TP), false positive (FP), false negative (FN) or true negative (TN). The capabilities of each SVMs were analysed with several statistical measures including:

Sensitivity Sensitivity (S_n) is calculated by Equation 1.7.2, also known as the True Positive Rate. The values for S_n range between 0 and 1 where a value of 1 correctly identifies all positive examples.

Specificity Specificity is calculated Equation 2.2.25.

$$S_p = \frac{TN}{N} \equiv \frac{TN}{TN + FP}$$

Equation 2.2.25: Specificity

The values for S_p range between 0 and 1 where a value of 1 correctly identifies all negative examples.

Matthews Correlation Coefficient The Matthews Correlation Coefficient is calculated by Equation 2.2.26.

$$MCC = \frac{((TP \times TN) - (FP \times FN))}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

Equation 2.2.26: Matthews Correlation Coefficient

The values range between -1 and 1, where 1 is a perfect prediction, 0 is a random prediction and -1 is an inverse prediction.

2.2.5 Updates to the Orthology and Transitive Modules

This section describes modules that have been imported from the PIPs 1 framework, but have had only minor alterations and have not been the main focus of study for the thesis. Described are the Orthology module and the Transitive module and the data that is used.

Orthology Module

The purpose of the orthology module is to identify interactions between proteins in other species that have an evolutionary relationship to proteins that are present within the human proteome. For two proteins (A and B) that are known to interact in one organism, if there are two orthologous proteins in another species (A' and B') that are also known to interact then this is known as an interolog (Walhout et al., 2000). The orthology module predicts two proteins in human that have interacting orthologous proteins in another species to be more likely to interact than to not interact (Scott and Barton, 2007).

The InParanoid database was used to identify orthologues (Berglund et al., 2008). InParanoid clusters the proteins into groups of homology and outputs a score representing how likely two proteins are of being orthologs. The score is based on how

similar proteins are to the two most homologous proteins within the cluster.

The evidence for interactions in different organisms is derived from 3 databases, DIP (Salwinski et al., 2004), HPRD (Keshava Prasad et al., 2009) and IntAct (Aranda et al., 2009). These databases use the PSIMI annotation hierarchy (Kerrien et al., 2007b) to describe the methods employed to identify protein-protein interactions. To filter protein-protein interactions, the highest common node (MI:0045) is used to identify that they have experimental evidence available. The MI:0045 node does encompass the nodes for genetic inference (MI:0254) and post translational interference (MI:0255), but there are no interactions present within the data that have been loaded into the database from DIP or IntAct with these evidence codes. Interactions that are labelled as genetically interacting are not included because the aim of the PIPs predictor is protein-protein interactions and two proteins that genetically interact do not always physically interact. Post translational interference refers to methods that interfere with the level of protein production, this too also does not indicate that two proteins interact. The HPRD use the old PSIMI codes for in vivo, in vitro and yeast-2-hybrid, which have become obsolete, but the codes remain within the database to accommodate for this.

The code base for the module has been changed so that known interactions and orthologs/paralogs are held in memory as opposed to being queried on demand. This has reduced the run time for the training and testing of the module from hours to minutes allowing for faster development of the code and testing new parameters.

Transitive Module

The transitive module was included in the first PIPs predictor to identify protein pairs that are likely to interact based on the local topology of the interaction network predicted by the Expression, Orthology and Combined modules. The transitive module requires predictions to be made by the Expression, Orthology and Combined modules before it can be trained and make predictions. In PIPs 1 (Scott and Barton, 2007), while making the final set of predictions the predictor had to be run in series, however, to complete the task in a tractable time, the protein pairs were batched. Due to the batching process the score for transitive proteins had to be pre-calculated during the training process, as a result these were the only pairs that could then be used in the final predictor for the transitive module. The effect of using the likelihood ratios from the training set is that it does not provide complete coverage of all potential protein pairs and therefore favours those present during training.

For the PIPs 2 framework, Expression, Orthology and Combined modules were run in parallel in their entirety. The results were integrated and then the protein pairs with a likelihood ratio ≥ 10 were saved for use by the Transitive module. This meant that the predictions made by the Transitive module could utilise all of the potential interactions calculated by Expression, Orthology and Combined modules.

2.3 Results

2.3.1 Analysis of Annotated Gene Ontology Terms as Part of the Combined Module

To investigate the inclusion of Gene Ontology annotations within the Combined module, the interaction likelihood ratio scores were calculated separately for each branch of the GO as summarised in table 2.2. Due to the likelihood ratio scores for each of the individual GO term modules being substantially less than 400 (the minimum likelihood ratio to obtain a posterior odds ratio of 1, see Section 1.6 for further details), the decision was taken to include the GO module as part of the Combined module.

Table 2.2: Likelihood Ratios calculated using each of the three different branches of the Gene Ontology: Cellular Compartment (C); Molecular Function (F); Biological Process (B)

Cut Off Point	C	F	B
< 0.2	0.65	0.68	0.67
$0.2 \leq x < 0.6$	3.33	4.5	5.65
≥ 0.6	4.28	5.1	6.38

Integration of the Combined and GO modules was performed using a full Bayesian method by calculating the maximum LR score for all possible combinations of the original 3 features from the Combined module (Domain co-occurrence, Localisation and Post Translational Modifications) and the 3 features from the GO (Cellular Compartment, Molecular Function and Biological Process). A BIC score (Equation 2.2.7) was then calculated based on the maximum likelihood ratio score and the number of bins involved within the module. All of the calculations used the same positive and negative datasets (size ratio 1:10). The most viable combinations of

features for a combined module are listed in Table 2.3.

Table 2.3: This table shows all potential combinations of features that can be considered by the Combined module: Co-occurrence of domains (D); Co-localisation (L); Co-occurrence of post translational modifications (P); GO Cellular Compartment (C); GO Molecular Function (F); GO Biological Process (B). The table is ordered in descending order of BIC score.

DLP Features	GO Features	Number of Bins	Max. LR Value	BIC Score
DLP	CFB	2160	2400	31827.65
DLP		80	3000	1163.37
DL	C	60	1327.27	870.15
DP	C	60	1712.5	869.64
DL	F	60	2000	869.33
DP	F	60	2100	869.23
DL	B	60	1525	869.87
DP	B	60	1800	869.54
LP	C	48	400	695.64
LP	F	48	544.44	695.03
LP	B	48	3900	691.09
D	CF	45	715.39	650.25
D	CB	45	462.5	651.13
D	FB	45	487.5	651.02
L	CF	36	28.58	524.01
L	CB	36	52.38	522.8
L	FB	36	50.71	522.87
P	CF	36	740	517.51
P	CB	36	700	517.62
P	FB	36	666.67	517.72
	CFB	27	24.7	382.47

The three combinations of features selected for further analysis were:

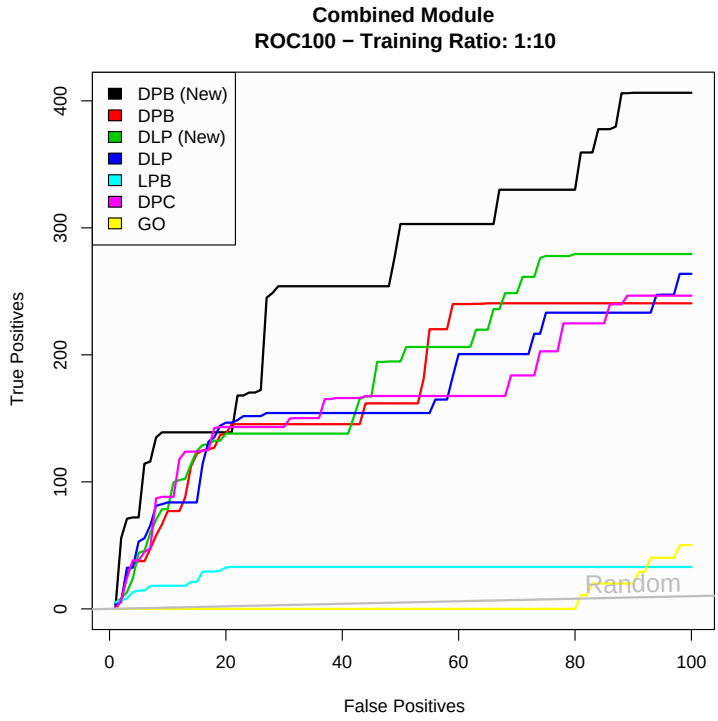
1. Domain Co-occurrence, Post Translational Modifications and Cellular Compartment (DPC),

2. Domain Co-occurrence, Post Translational Modifications and Biological Process (DPB),

3. Localisation, Post Translational Modifications and Biological Process (LPB).

The reason for selecting DPC, DPB and LPB was that they had the three highest LR values (other than DLP) and BIC scores less than the DLP combination as

Figure 2.4: ROC100 curves, a plot of the top 100 false positive predictions against the number of true positive predictions. The (New) data refers to the the modules that have been trained with the HPRD 2007 release of the database as the positive training set.



used in the Combined module of PIPs 1. These three were also selected based on time limitations and the use of Biological Process rather than Molecular function seemed a more logical selection. Models were then trained with these new Combined modules and ROC curves were calculated to determine the effectiveness of the predictor. Table 2.4 shows the area under the ROC100 curves (AUC) values for the new Combined modules trained using the PIPs 1 datasets (Scott and Barton, 2007). When compared using a ROC100 plot (Figure 2.4) the predictive capability of LPB to predict positive results with high likelihood ratio values is limited. DPB was chosen for further analysis over DPC due to consistently choosing more positives than negatives at higher likelihood ratio values.

Table 2.4: Partial ROC100 area under curves for Figure 1 for predictors calculated using the training sets used by the PIPs 1 predictor.

Module	ROC100 AUC
DLP (Original Module)	16900.0
DPB	17700.0
DPC	16600.0
LPB	3000.0

The protein-protein interactions in the PIPs 1 database were updated to come in line with the current release of the HPRD (version 7) and version 3.40 of the IPI database. The DLP and DPB Combined modules were trained using the new data to determine if there is an increase in accuracy of the predictions that were made. Table 2.4 summarises the partial ROC AUC results. For both modules the ROC100 curve (Figure 2.4) shows there is an increase in the number of positive predictions made by both modules, but that the DPB module predicts over 400 positive predictions for the first 100 false positive predictions in comparison to less than 300 for the DLP module trained on the HPRD 2007 dataset.

2.3.2 Accuracy of the Cluster Module and Clustering of the Predicted Interactome

This Section presents the results of clustering of protein interactions networks, both of known and predicted interactions and teh creation of a Clustering module for inclusion within the PIPs predictor.

Clustering Known Interactions

Known sets of complexes were used to determine the capability of the MCL algorithm. The known set of biological complexes was downloaded from the HPRD

complex datasets and then filtered to create a "True Complex" set. Figure 2.5 shows what is meant by a "True Complex"; where there are sufficient known protein-protein interactions in the HPRD that each protein is connected in a single complex. For example, Figure 2.5A is a single cluster as there exists a path connecting all nodes and therefore is identified as a True Complex. Figure 2.5B forms two clusters with the same number of edges and would be defined to not be a True Complex. Out of the 1060 complexes within the HPRD, only 977 are present in the True Complex set.

When clustering the network of proteins present in the True Complex set, weighted by their LR_{EOC} values, MCL generates a list of 272 clusters. This is substantially less than the 997 true complexes within the HPRD. The reason for this is that within the network a single protein is represented by a single node. When clustering, MCL can only assign a protein to a single cluster, but in nature that same protein can

Figure 2.5: Network diagram of two complexes with 5 protein-protein interactions as labelled within the HPRD. Each node is representative of a protein and the edges represent interaction. (A) Represents a True Complex where all the proteins form a single linked group. (B) Represents a set of proteins labelled as a complex, but there are not sufficient interactions to form a single linked group.

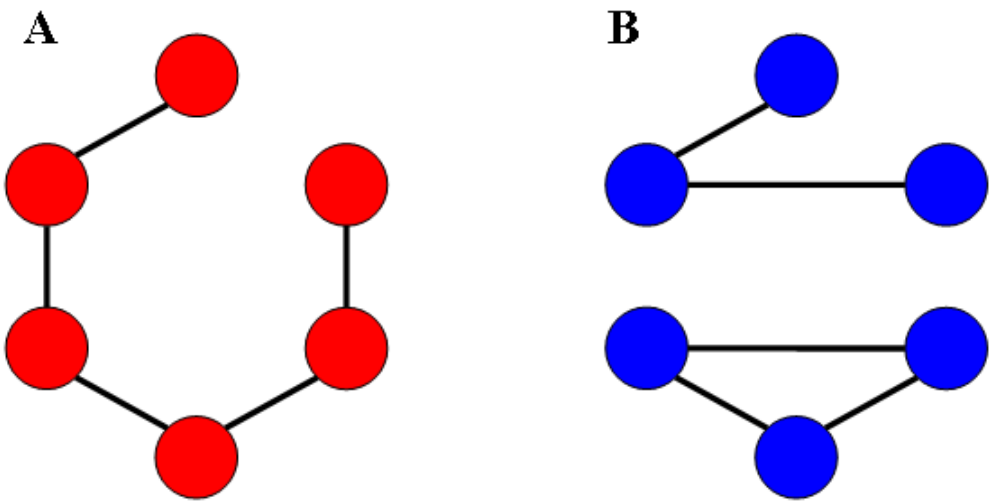
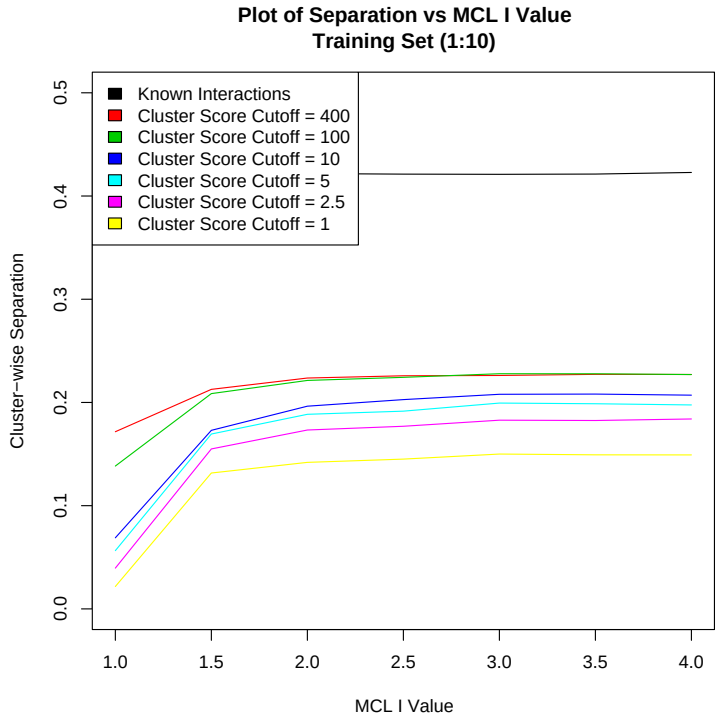
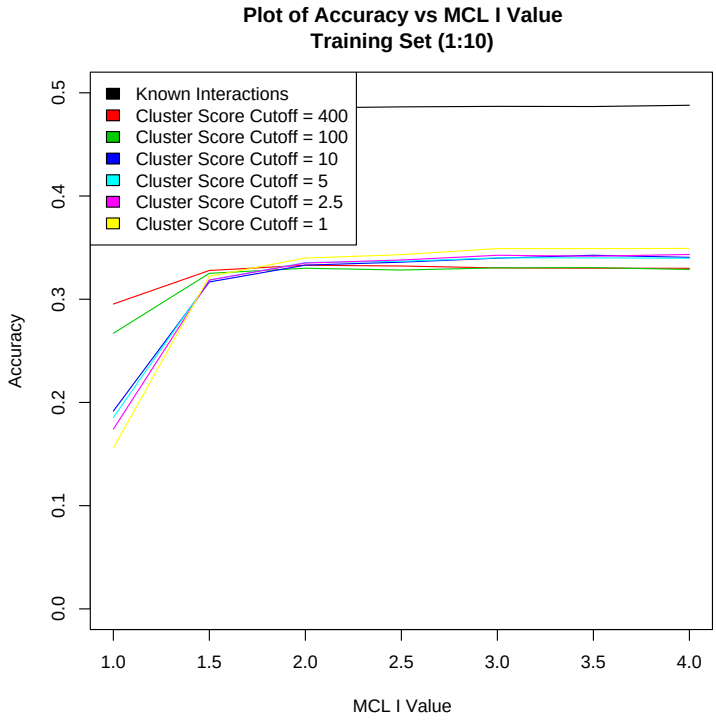


Figure 2.6: Comparison of the effect of varying the MCL I value (inflation argument) against the separation of the clusters when they are compared to known complexes. Values are also shown for the separation when clustering training set data with various likelihood ratio threshold levels. The Cluster Score Cutoff is the LR_{EOC} threshold point used to select protein pairs for clustering



exist within multiple complexes. For a protein that is present in two complexes, the proteins of both complexes are likely to be clustered together. This is shown in Figure 2.6 where there is a reduction in the separation score, which is less than 0.5 (black line). This mixing of multiple complexes also results in a reduction of the accuracy (Accuracy < 0.5), as shown in Figure 2.7 (black line). Both Figures 2.6 and 2.7 are plotted against the I arguments value for MCL. The I value represents the inflation value and is required for regulating the granularity of the clusters.

Figure 2.7: Comparison of the effect of varying the MCL I value (inflation argument) against the accuracy of the clusters when they are compared to known complexes. Values are also shown for the accuracy when clustering training set data with various likelihood ratio threshold levels. The Cluster Score Cutoff is the LR_{EOC} threshold point used to select protein pairs for clustering

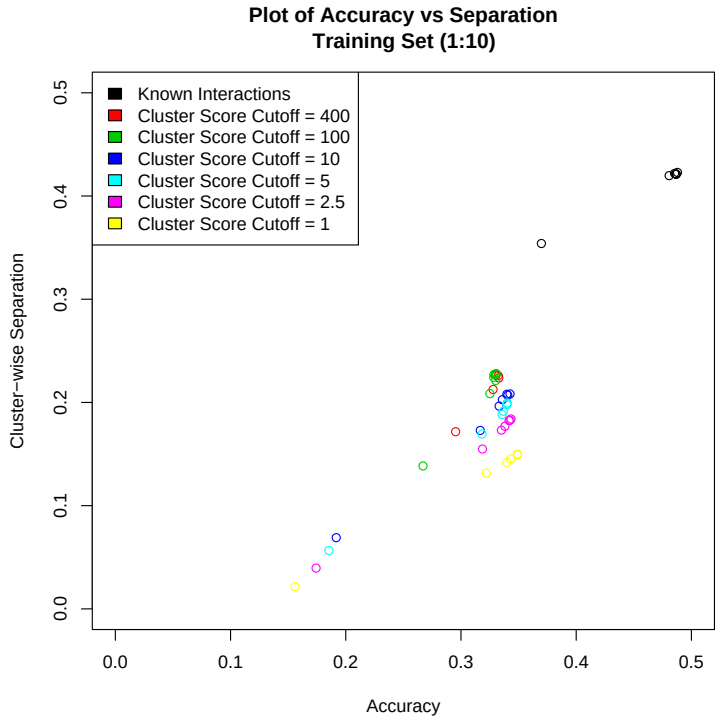


Parameter Selection

MCL I Value The default I value for MCL is 2.0. As shown in Figure 2.6 and Figure 2.7 there is a plateau in the accuracy and the separation compared to increase in the I value at 2.0. An I value of 2.0 has therefore been selected for all further clustering analysis.

Clustering Likelihood Ratio Threshold Figure 2.8 shows that increasing the threshold for the protein pairs that are clustered results in an increase in the accuracy and separation of the clustering. However, there is a reduction in the coverage of proteins. Table 2.5 shows the coverage of the proteome dependent on the set LR_{EOC}

Figure 2.8: Plot of the Accuracy verses the Separation scores from Figures 2.6 and 2.7. Each point is at a different I value, where increasing the I value tends to improve both the Accuracy and the separation. The Cluster Score Cutoff is the LR_{EOC} threshold point used to select protein pairs for clustering



threshold during training.

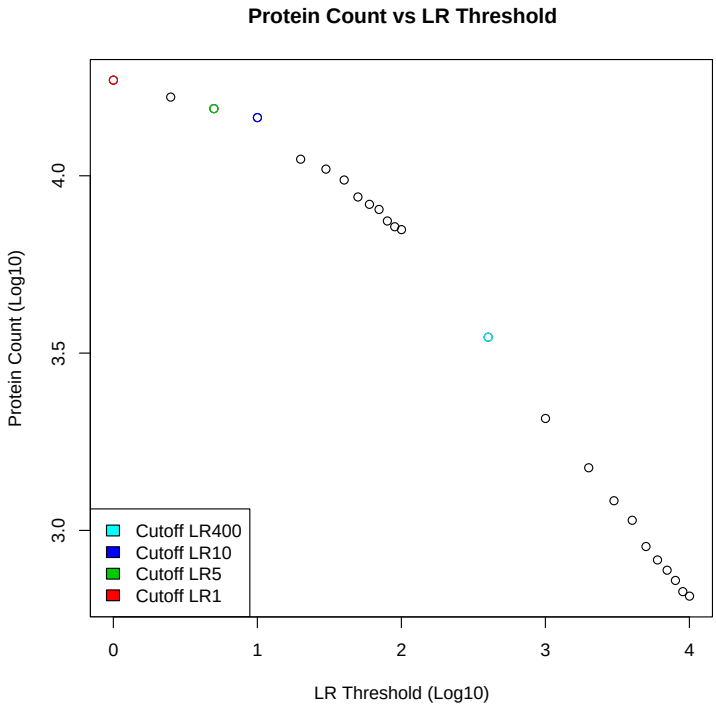
Raising the threshold therefore reduces the coverage of the proteome and as a result reduces the number of potential predictions that can be made. Figure 2.9 shows the Log-Log plot of the number of proteins (coverage) against various LR_{EOC} threshold levels. Once the LR_{EOC} threshold goes above 10 (blue), the coverage of the proteome reduces drastically. With an MCL I value fixed at 2.0, LR_{EOC} thresholds of 5 and 10 both have a similar accuracy (0.26 and 0.25 respectively) and separation (0.109 and 0.101) statistics. Therefore, an LR_{EOC} threshold of 5 was selected to maximise the coverage of the proteome. To determine whether to select an LR_{EOC} threshold of 5 or 2.5 it is beneficial to look at the calculated likelihood

Table 2.5: Table shows the count of the number of proteins involved in interactions that have an LR_{EOC} value above of set thresholds. Figure 2.9 shows a plot over the threshold range of 1 to 10000.

LR_{EOC} Value	Number of Proteins
≥ 1.0	18603
≥ 5.0	15470
≥ 10.0	14598
≥ 100.0	7047
≥ 400.0	3507

ratios. Table 2.6 shows that a threshold of 5 results in greater likelihood ratios for each of the classification bins over that of an LR_{EOC} threshold of 2.5.

Figure 2.9: Log-Log plot of the coverage of proteins with different likelihood ratio threshold levels with highlighted points of interest.



Final Module Analysis

Figure 2.10 shows the ROC100 plot for the PIPs predictor without a network analysis module (EOC, blue line) in comparison to when it is combined with a network

Table 2.6: Calculated likelihood ratios generated dependent on LR_{EOC} threshold selected for generating the network for clustering. C_x is the score assigned to a cluster.

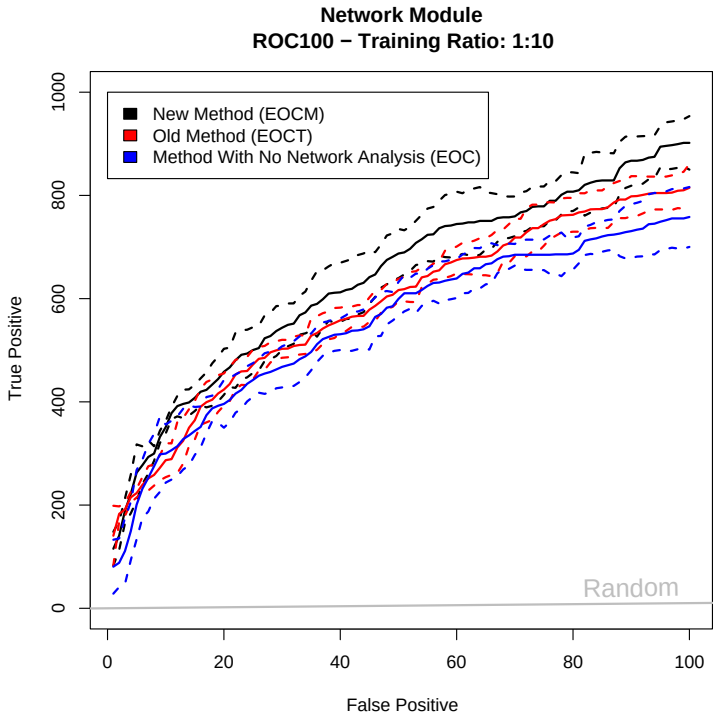
Cluster Score C_x	LR_{EOC} Threshold 1	2.5	5	10
Separate Clusters	0.9	0.93	0.93	0.94
$C_x \leq 50$	4.77	7.69	11.08	15.26
$50 < C_x \leq 100$	4.08	19.77	27.12	39.89
$100 < C_x \leq 400$	2.81	21.13	30.15	43.69
$400 < C_x \leq 1000$	3.04	24.56	42.81	60.35
$C_x > 1000$	3.33	19.88	35.62	69.07

analysis module (black and red lines). Figure 2.10 shows that including a network analysis module increases the accuracy of the protein-protein interaction predictions. After the first 100 false positive predictions, EOCT predicted 814.8 ± 45.9 true positives and EOCM predicted 901.8 ± 58.1.

Clustering and Transitive Modules in the PIPs 2 Predictor

The clustering and transitive analysis modules can not both be included within the PIPs 2 predictor. The reason for this is that the naïve Bayesian model requires that all modules are independent. The likelihood ratios assigned to protein pairs within the test sets by the clustering and transitive modules have a Pearson's correlation coefficient of 0.365. As a result they are not statistically different for both to be included within the final predictor. It would be possible to group the two modules together as part of a full Bayesian module to calculate a likelihood ratio in a similar method to the Combined module.

Figure 2.10: ROC 100 plot comparing the PIPs framework based on different module compositions: Expression Module (E), Orthology (O), Combined (C), Transitive (T), MCL Clustering (M). EOC represents the PIPs predictive framework without a network analysis module. The dotted lines indicate 1 standard deviation during cross validation.



2.3.3 Accuracy of the Gene Expression Module

This Section describes the investigation of different methods of correlation to measure co-expression of genes and the use of different expression datasets to improve the accuracy of the Expression module.

Correlation of Gene Coexpression

Figure 2.11 compares in a pair wise manner the different measures of correlations of gene co-expression values for the A-AFFY-47 combined dataset. The largest difference is between Pearson's and the more robust methods such as Spearman's, Kendall and Hardin.

The most similar co-expression values are between Spearman's and Kendall. This is because of the similarity in the way that each of the equations handles the data. Both are ranked correlation methods that calculate the difference in ranks. As a result it is not a surprise to observe similar values for the two methods.

Even though the Hardin equation results in the most robust correlation measure, the code is only available within R and is computationally very expensive. As a result only 1000 correlations can be calculated per minute. This would require 3 days to calculate a full set of gene co-expression correlation values. As a compromise, Spearman's calculates similar correlation values (Figure 2.11e), but can calculate substantially more correlations per minute. This is because the Hardin et al. (2007) method iteratively refines the parameters for each correlation, whereas there is no refinement step required within Spearman's. The Hardin et al. (2007) can be made to run faster by improving the code or moving it to a compiled language such as Java or C, but such refinements are beyond the scope of this investigation.

Figure 2.11: Comparison of different correlation methods. The figures a-f show the comparison between the different correlation methods. On plot (a) the characters A-F refer to Figure 2.12 to highlight the difference between the correlation calculated by Pearsons or Spearmans measures.

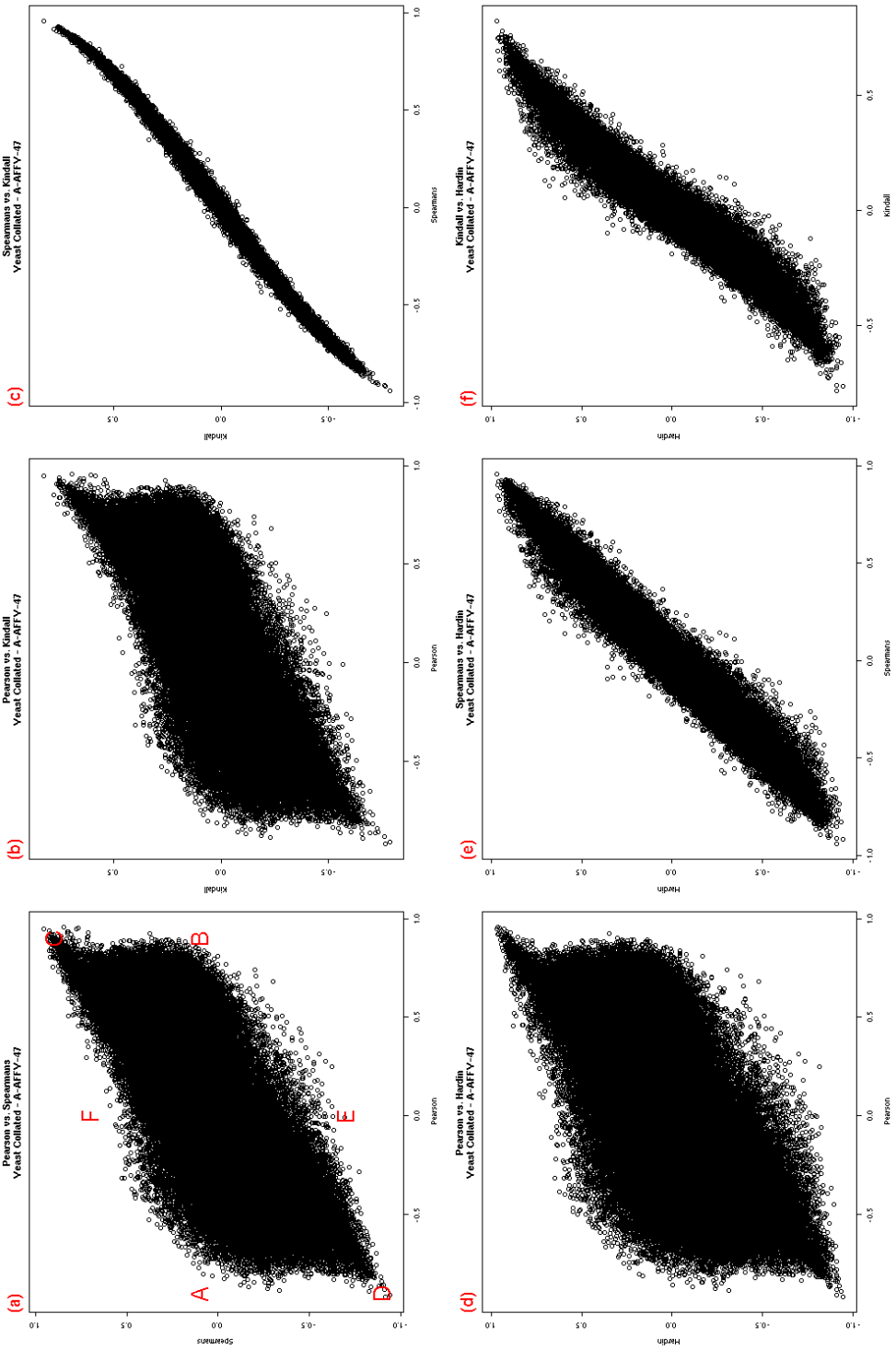


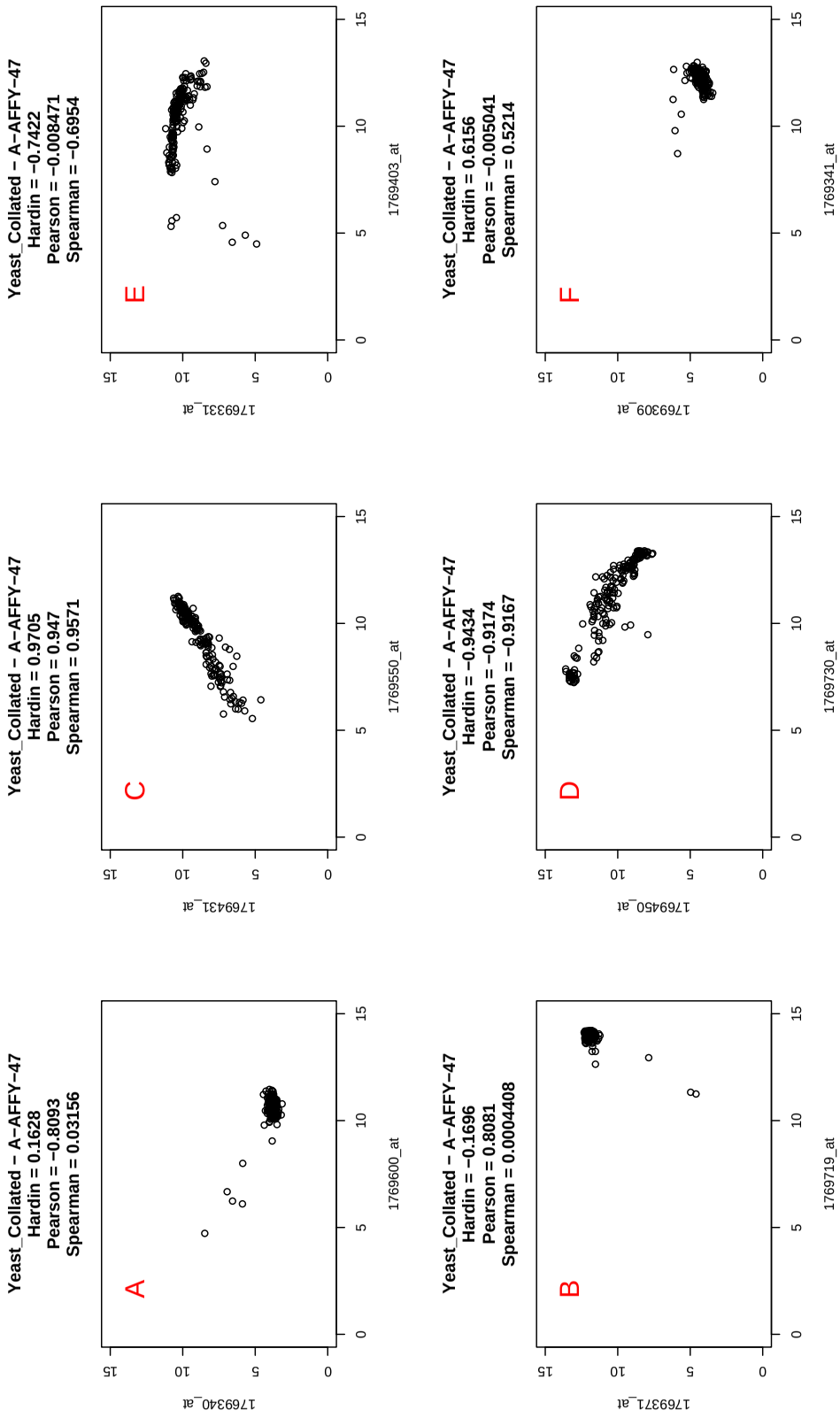
Figure 2.12 shows example gene pairs from the regions labelled A-E on Figure 2.11a. Figures 2.12C and 2.12D show gene expression levels that are correlated with few outliers that do not drastically deviate. As a result both Spearman's and Pearson's correlations are similar (Figure 2.12C: 0.957 and 0.941; Figure 2.12D: -0.917 and -0.917 respectively). It is when there are outliers that deviate drastically from the rest of the gene expressions that there is a larger effect on the correlation values. For example in Figure 2.12A Spearman's and Hardin have a low correlation value, yet Pearson's identifies the outliers as part of the trend therefore indicating that there is a strong positive correlation. This is also observed for negative correlations (Figure 2.12B). To a lesser extent Figures 2.12E and 2.12F show examples of gene expression levels where Spearman's has given a mildly positive correlation, but Pearson's has found no correlation.

It is very important to select the right method to calculate correlation. Selection will depend on whether it is best to include all points from the dataset, or if by including all points this will introduce noise that will pollute the final results. For the Expression module both Pearson's and Spearman's measure of correlation were calculated and the final decision was be based on the most accurate predictor as measured by ROC 100 plots.

Comparison of Human Gene Expression Training Datasets

The E-TABM-145 (GEO596) dataset was used both by Scott and Barton (2007) and in this study to analyse improvements due to increase in the number of proteins within the IPI (Kersey et al., 2004) and whether improving the calculation for correlation would have an effect on the likelihood of interaction.

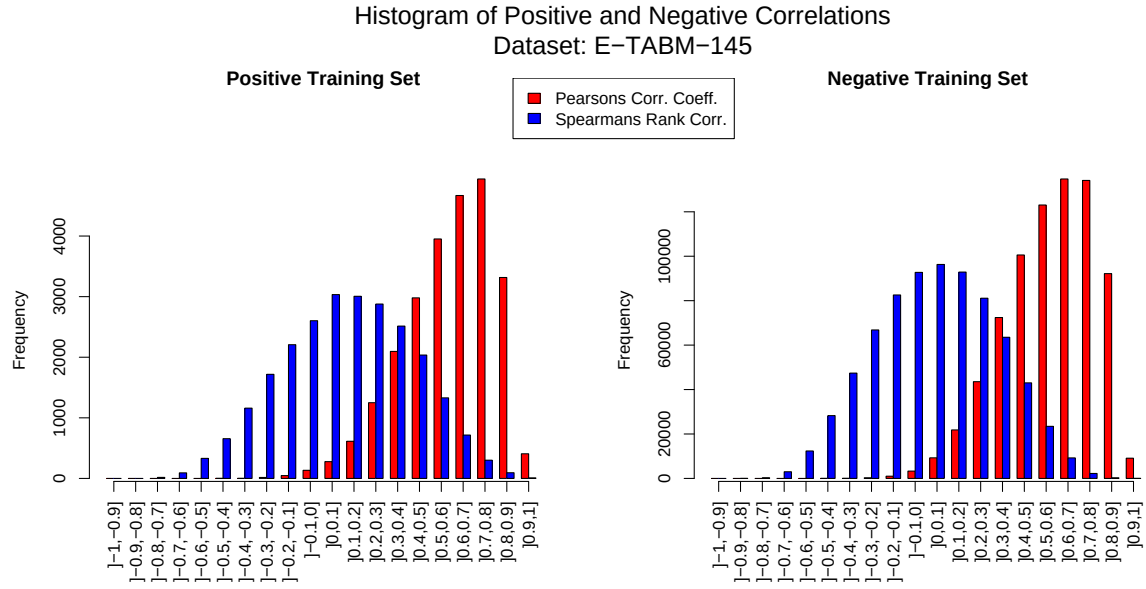
Figure 2.12: Plots of selected genes to emphasis the difference in correlation as measured by Pearson' and Spearman's correlation coefficients.



The first major difference between the dataset used by Scott and Barton and the same dataset that I have used is the coverage. Due to changes in source databases, such as the IPI database (Kersey et al., 2004), the number of matching proteins between the A-AFFY-33 chip was increased from 10,642 (Scott and Barton, 2007) to 13,639.

Figure 2.13 compares the distribution of the correlation values of the positive and negative training sets using the Pearson's and Spearman's rank correlations. The graphs highlight the difference between using robust versus a non-robust measure to calculate correlation. Pearson's correlation shows that there is a bias towards allocating a positive correlation. However Spearman's has a more normal distribution for both the proteins pairs within the positive and negative training sets.

Figure 2.13: Plots of the distribution of gene co-expression values for the training sets (Positive left, Negative right). Red representing correlations calculated using Pearson's correlation and blue representing Spearman's rank correlation.



To compare the predictive accuracy of the expression datasets, Figure 2.14 shows

the ROC100 curves for predictors built using the E-TABM-145 dataset with either Pearson's or Spearman's correlation. What the graph shows is that using the Spearman's measure of correlation (black line) outperforms the predictor that uses Pearson's correlation (green line). However, the predictive capability of the dataset is not very strong. The ROC 100 curves (Figure 2.15) show that for the first 100 false positive predictions, the number of true positive predictions is less than 100.

Figure 2.14: ROC100 plot of the comparison of using the E-TABM-145 dataset used in PIPs 1 with filtered and unfiltered probes and changing the measure of correlation. The grey line represents random selection

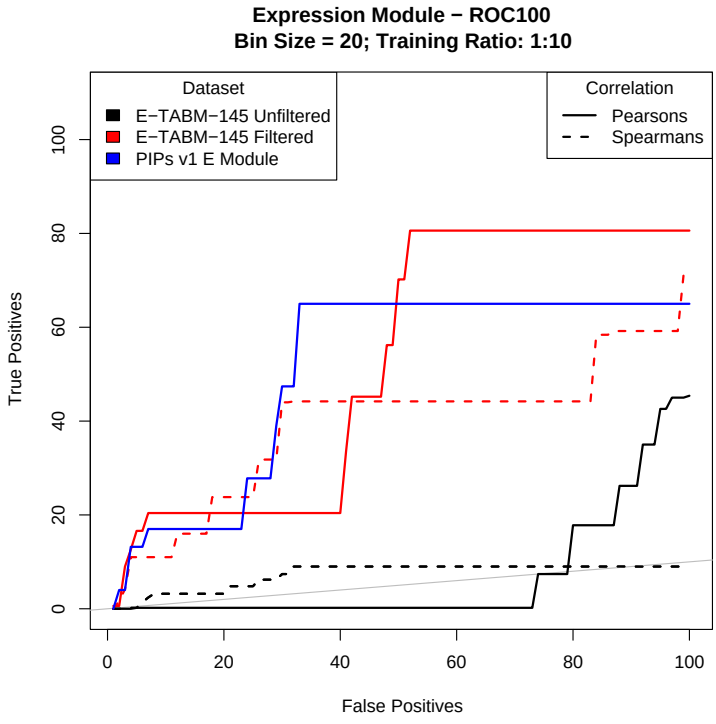
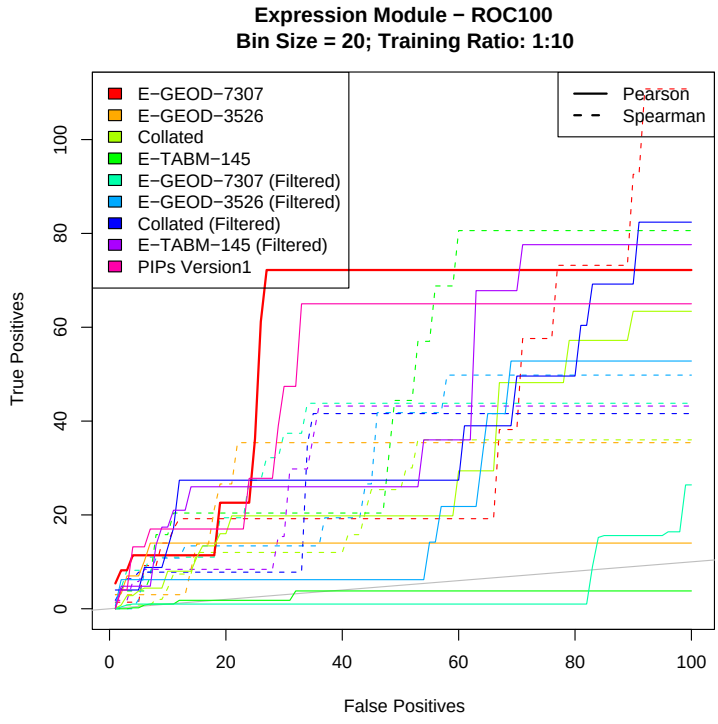


Figure 2.15: ROC100 plot of all potential Gene Expression sets that could be used within the Expression module. The bold red line indicates the select dataset and correlation measure. The grey line represents random selection



Multiple Proteins per Probe

Based on the index files supplied by AffyMetrix, some of the probes on the A-AFFY-33 and A-AFFY-44 chips map to multiple proteins. The A-AFFY-33 chip has probes for 13639 proteins; however 3831 of the probes match multiple proteins. This problem is also present in the A-AFFY-44 where there are probes that link to regions that encode 18334 proteins, but there are 6517 probes that match to more than 1 protein.

There are 1970 probes on the A-AFFY-33 and 5891 on the A-AFFY-44 chip that have multiple proteins associated to that probe, but where the associated proteins also have a second matching probe that they uniquely associate to.

Figure 2.16 shows probe 210825_s_at from the A-AFFY-33 chip matches 3 different proteins, where each protein matches a unique probe as well. The average expression for the other proteins is plotted on the graph. Some of the proteins have expression profiles that have a high correlation to the multi-matching probe. Other proteins have little to no match.

Figure 2.17 shows a histogram of the correlation between the average expression profile for probes that uniquely match a protein to a probe that matches two proteins. The profile suggests that the profile of the multimatch probe rarely matches the average expression profile of the probes that uniquely match the protein. However, as shown in Figure 2.15, filtering out these probes does not add to the predictive capability of the module and as such, it was decided to consider all probes within the predictor.

Figure 2.16: The plot shows the expression level detected by 4 probes. The probe 210825_s_at matches 3 proteins (IPI00019761, IPI00219446 and IPI00219682). Each protein has a matching unique probes (red, blue and green) where the average of the unique probes for each protein are plotted.

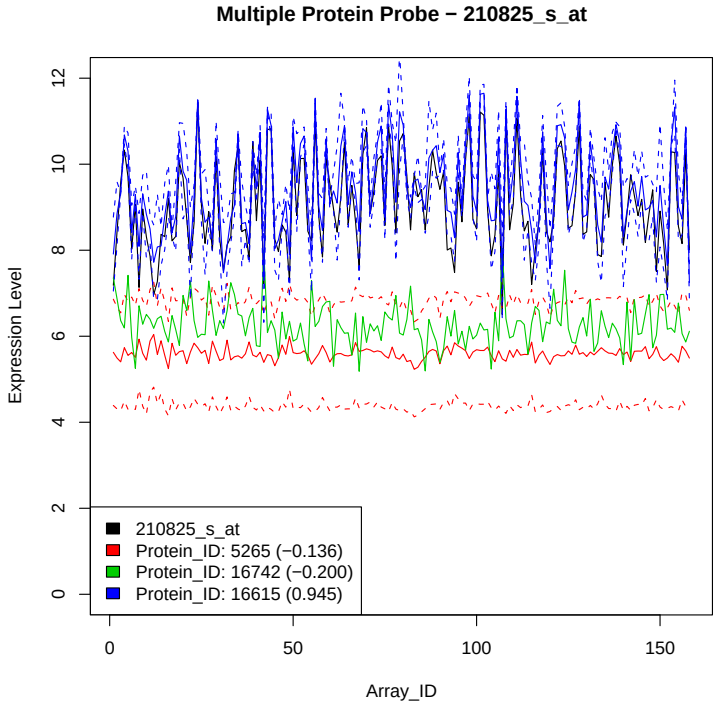
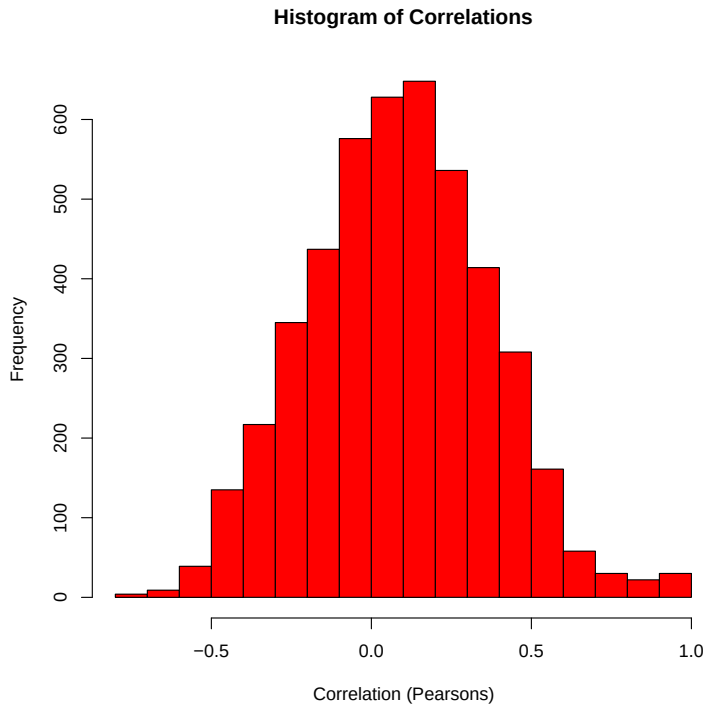


Figure 2.17: Histogram of gene co-expression correlations.



2.3.4 Predictive Capability of the Sequence Module

The results for each of the individual SVMs (as described in Section 2.2.4) are summarised in Tables 2.7-2.11, showing only the top 20 results as ordered by Matthews Correlations.

Tripeptide Secondary Structure Motif

Because the results for the standardised and scaled training sets are almost identical the scaled values were used for making predictions based on Jpred motifs (data not shown).

Tables 2.7 and 2.8 are the top 20 SVMs that were trained on scaled data and ranked by the Matthews Correlation Coefficient. Table 2.7 shows the results of training SVMs with the full Jpred motif feature set. The SVMs in Table 2.8 were trained with the reduced Jpred motif feature set. As is evident from Tables 2.7 and 2.8 the predictors are weakly predictive with Matthews correlation coefficients less than 0.3.

Full Tripeptide Motifs Optimal SVM parameters for SVMs were selected by filtering for Matthews' correlation coefficients $\geq$0.18. When maximising for Matthews correlation two potential SVM models, both utilising the Laplace kernel, have the highest correlation. The cost function selected was 1.0 as it had a greater specificity for a similar sensitivity (0.613 and 0.57 respectively) over the Laplace kernel with a cost function of 0.5 (0.591 and 0.597 respectively).

Table 2.7: Jpred Motifs (27) results for the top 20 SVMs ordered by their Matthews Correlations trained with 1000 positive and 1000 negative examples. The TP, FP, FN and TN values are from the classification of the 33094 positive and 33094 negative test example protein pairs. (MCC: Matthews Correlation Coefficient)

Kernel	Correlation	Cost	No. Vectors	TP	FP	FN	TN	Sensitivity	Specificity	MCC
laplacedot	Spearman	0.1	1776	22842	16050	10252	17044	0.69	0.515	0.20846
laplacedot	Pearson	0.1	1779	22846	16060	10248	17034	0.69	0.515	0.20829
rbfdot	Pearson	0.1	1583	22026	15263	11068	17831	0.666	0.539	0.20602
rbfdot	Spearman	0.1	1584	22049	15289	11045	17805	0.666	0.538	0.20597
laplacedot	Spearman	0.5	1579	19764	13537	13330	19557	0.597	0.591	0.18816
laplacedot	Pearson	0.5	1576	19645	13455	13449	19639	0.594	0.593	0.18704
vanilladot	Spearman	20	1331	16441	10438	16653	22656	0.497	0.685	0.18468
vanilladot	Pearson	20	1330	16449	10446	16645	22648	0.497	0.684	0.18466
polydot	Spearman	20	1330	16441	10440	16653	22654	0.497	0.685	0.18461
polydot	Pearson	20	1328	16444	10444	16650	22650	0.497	0.684	0.18458
vanilladot	Pearson	15	1329	16563	10564	16531	22530	0.5	0.681	0.18429
polydot	Pearson	15	1330	16566	10567	16528	22527	0.501	0.681	0.18429
vanilladot	Spearman	15	1330	16561	10568	16533	22526	0.5	0.681	0.18411
polydot	Spearman	15	1329	16561	10569	16533	22525	0.5	0.681	0.18407
laplacedot	Spearman	1	1526	18856	12794	14238	20300	0.57	0.613	0.18335
polydot	Spearman	10	1329	16573	10606	16521	22488	0.501	0.68	0.18326
vanilladot	Spearman	10	1328	16573	10608	16521	22486	0.501	0.679	0.18319
laplacedot	Pearson	1	1524	18875	12820	14219	20274	0.57	0.613	0.18313
vanilladot	Pearson	10	1327	16573	10611	16521	22483	0.501	0.679	0.1831
polydot	Pearson	10	1326	16574	10613	16520	22481	0.501	0.679	0.18306

Table 2.8: Jpred Motifs (10) results for the top 20 SVMs ordered by their Matthews Correlations trained with 1000 positive and 1000 negative examples. The TP, FP, FN and TN values are from the classification of the 33094 positive and 33094 negative test example protein pairs.

Kernel	Correlation	Cost	No. Vectors	TP	FP	FN	TN	Sensitivity	Specificity	MCC
rbfdot	Spearman	0.1	1526	20848	13707	12246	19387	0.630	0.586	0.216
rbfdot	Pearson	0.1	1528	20690	13560	12404	19534	0.625	0.590	0.216
laplacedot	Pearson	0.1	1631	22832	15892	10262	17202	0.690	0.520	0.213
laplacedot	Spearman	0.1	1647	22832	15897	10262	17197	0.690	0.520	0.213
laplacedot	Spearman	0.5	1523	19628	12653	13466	20441	0.593	0.618	0.211
laplacedot	Pearson	0.5	1518	19515	12542	13579	20552	0.590	0.621	0.211
laplacedot	Pearson	1	1482	18119	11231	14975	21863	0.548	0.661	0.209
laplacedot	Spearman	1	1486	18184	11317	14910	21777	0.549	0.658	0.209
polydot	Spearman	20	1355	16806	10289	16288	22805	0.508	0.689	0.200
polydot	Pearson	20	1354	16803	10287	16291	22807	0.508	0.689	0.200
polydot	Pearson	15	1351	16804	10290	16290	22804	0.508	0.689	0.200
vanilladot	Spearman	20	1352	16800	10287	16294	22807	0.508	0.689	0.200
vanilladot	Pearson	20	1352	16803	10291	16291	22803	0.508	0.689	0.200
polydot	Spearman	15	1353	16809	10298	16285	22796	0.508	0.689	0.200
vanilladot	Spearman	10	1350	16810	10299	16284	22795	0.508	0.689	0.200
vanilladot	Spearman	15	1351	16805	10295	16289	22799	0.508	0.689	0.200
vanilladot	Pearson	15	1353	16806	10296	16288	22798	0.508	0.689	0.200
polydot	Pearson	10	1350	16809	10299	16285	22795	0.508	0.689	0.200
polydot	Spearman	10	1350	16809	10299	16285	22795	0.508	0.689	0.200
vanilladot	Pearson	10	1349	16810	10302	16284	22792	0.508	0.689	0.200

Table 2.9: Sequence Motifs (343) results for the top 20 SVMs ordered by their Matthews Correlations trained with 1000 positive and 1000 negative examples. The TP, FP, FN and TN values are from the classification of the 33094 positive and 33094 negative test example protein pairs.

Kernel	Correlation	Cost	No. Vectors	TP	FP	FN	TN	Sensitivity	Specificity	MCC
laplacedot	Spearman	0.5	1905	7879	5829	25215	27265	0.238	0.824	0.076
laplacedot	Spearman	0.1	2000	13127	10783	19967	22311	0.397	0.674	0.074
rbfdot	Spearman	0.1	1819	3852	2448	29242	30646	0.116	0.926	0.072
laplacedot	Pearson	0.5	1917	8146	6197	24948	26897	0.246	0.813	0.071
laplacedot	Pearson	0.1	2000	12818	10679	20276	22415	0.387	0.677	0.068
rbfdot	Pearson	0.1	1826	3642	2431	29452	30663	0.110	0.927	0.063
laplacedot	Spearman	1	1808	1978	1332	31116	31762	0.060	0.960	0.045
laplacedot	Pearson	1	1809	1965	1413	31129	31681	0.059	0.957	0.038
tanhdot	Spearman	1	1002	17621	16657	15473	16437	0.532	0.497	0.029
tanhdot	Spearman	0.5	1017	17757	16795	15337	16299	0.537	0.493	0.029
tanhdot	Spearman	15	980	17611	16675	15483	16419	0.532	0.496	0.028
tanhdot	Spearman	10	981	17610	16676	15484	16418	0.532	0.496	0.028
tanhdot	Spearman	20	980	17608	16675	15486	16419	0.532	0.496	0.028
tanhdot	Spearman	5	981	17597	16667	15497	16427	0.532	0.496	0.028
tanhdot	Spearman	0.1	1118	17797	16903	15297	16191	0.538	0.489	0.027
tanhdot	Pearson	0.5	1017	17272	16415	15822	16679	0.522	0.504	0.026
tanhdot	Pearson	0.1	1098	17680	16840	15414	16254	0.534	0.491	0.025
tanhdot	Pearson	10	973	17049	16239	16045	16855	0.515	0.509	0.024
tanhdot	Pearson	5	973	17039	16229	16055	16865	0.515	0.510	0.024
tanhdot	Pearson	1	978	17035	16228	16059	16866	0.515	0.510	0.024

Table 2.10: Sequence Motifs (84) results for the top 20 SVMs ordered by their Matthews Correlations trained with 1000 positive and 1000 negative examples. The TP, FP, FN and TN values are from the classification of the 33094 positive and 33094 negative test example protein pairs.

Kernel	Correlation	Cost	No. Vectors	TP	FP	FN	TN	Sensitivity	Specificity	MCC
laplacedot	Spearman	1	1861	14739	9227	18355	23867	0.445	0.721	0.17328
laplacedot	Pearson	1	1868	14130	8824	18964	24270	0.427	0.733	0.16843
laplacedot	Spearman	5	1650	11050	6238	22044	26856	0.334	0.812	0.1655
laplacedot	Spearman	0.5	1944	17342	12090	15752	21004	0.524	0.635	0.15968
laplacedot	Pearson	5	1664	10793	6226	22301	26868	0.326	0.812	0.15788
rbfdot	Spearman	0.5	1670	10362	5918	22732	27176	0.313	0.821	0.15591
laplacedot	Spearman	0.1	2000	19467	14335	13627	18759	0.588	0.567	0.15511
laplacedot	Pearson	0.5	1945	16781	11703	16313	21391	0.507	0.646	0.15495
rbfdot	Pearson	0.5	1665	10238	5931	22856	27163	0.309	0.821	0.15145
laplacedot	Pearson	0.1	2000	19777	14791	13317	18303	0.598	0.553	0.15081
laplacedot	Spearman	10	1639	9555	5416	23539	27678	0.289	0.836	0.14947
rbfdot	Spearman	0.1	1865	12051	7606	21043	25488	0.364	0.77	0.14697
laplacedot	Pearson	10	1659	9421	5386	23673	27708	0.285	0.837	0.14629
rbfdot	Pearson	0.1	1870	11691	7376	21403	25718	0.353	0.777	0.14396
laplacedot	Spearman	15	1639	8710	4915	24384	28179	0.263	0.851	0.14181
laplacedot	Pearson	15	1655	8628	4942	24466	28152	0.261	0.851	0.13794
rbfdot	Spearman	1	1574	8852	5129	24242	27965	0.267	0.845	0.1378
rbfdot	Pearson	1	1588	9152	5417	23942	27677	0.277	0.836	0.1362
laplacedot	Spearman	20	1663	8299	4726	24795	28368	0.251	0.857	0.13578
laplacedot	Pearson	20	1672	8286	4789	24808	28305	0.25	0.855	0.1327

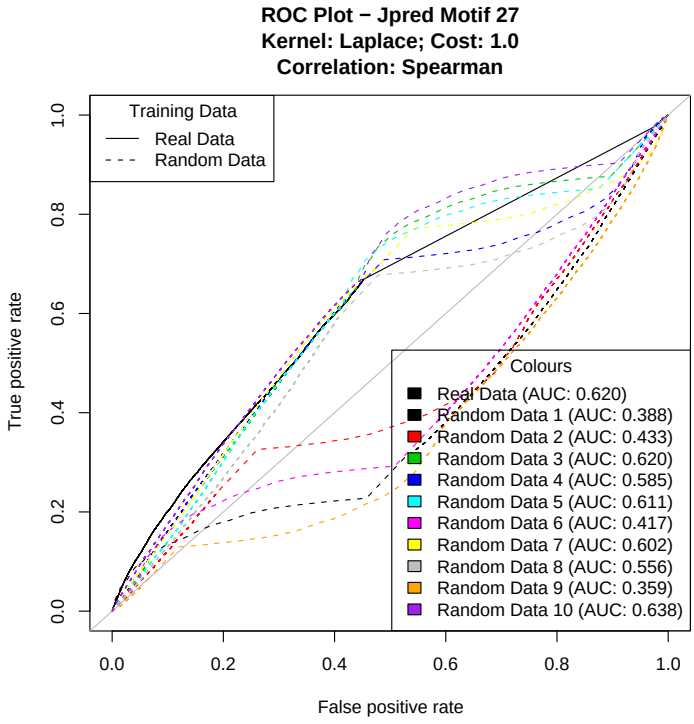
Table 2.11: Proportion results for the top 20 SVMs ordered by their Matthews Correlations trained with 1000 positive and 1000 negative examples. The TP, FP, FN and TN values are from the classification of the 33094 positive and 33094 negative test example protein pairs.

Kernel	Correlation	Cost	No. Vectors	TP	FP	FN	TN	Sensitivity	Specificity	MCC
laplacedot	Spearman	1	1670	17055	10284	16039	22810	0.515	0.689	0.20776
laplacedot	Pearson	1	1664	16955	10240	16139	22854	0.512	0.691	0.20621
laplacedot	Spearman	0.5	1717	17764	11054	15330	22040	0.537	0.666	0.20447
laplacedot	Pearson	0.5	1721	17684	11055	15410	22039	0.534	0.666	0.20207
laplacedot	Spearman	0.1	1879	20014	13369	13080	19725	0.605	0.596	0.2008
laplacedot	Pearson	0.1	1884	19894	13378	13200	19716	0.601	0.596	0.1969
rbfdot	Spearman	0.5	1587	14969	8952	18125	24142	0.452	0.729	0.18923
laplacedot	Spearman	5	1618	13946	8157	19148	24937	0.421	0.754	0.18545
rbfdot	Pearson	0.5	1583	14963	9079	18131	24015	0.452	0.726	0.18485
rbfdot	Spearman	1	1501	14215	8414	18879	24680	0.43	0.746	0.18477
laplacedot	Pearson	5	1616	13940	8216	19154	24878	0.421	0.752	0.18326
rbfdot	Pearson	1	1506	14353	8595	18741	24499	0.434	0.74	0.18279
rbfdot	Spearman	0.1	1755	15534	9799	17560	23295	0.469	0.704	0.17827
rbfdot	Pearson	0.1	1753	15506	9894	17588	23200	0.469	0.701	0.17436
laplacedot	Spearman	10	1656	12844	7542	20250	25552	0.388	0.772	0.17351
laplacedot	Pearson	10	1656	12763	7513	20331	25581	0.386	0.773	0.17207
laplacedot	Spearman	15	1701	12370	7296	20724	25798	0.374	0.78	0.16775
laplacedot	Pearson	15	1704	12283	7289	20811	25805	0.371	0.78	0.16533
laplacedot	Spearman	20	1738	12104	7260	20990	25834	0.366	0.781	0.16087
laplacedot	Pearson	20	1733	12061	7241	21033	25853	0.364	0.781	0.16022

Reduced Tripeptide Motifs Table 2.8 shows various parameters used to train SVMs and the corresponding Matthews correlation coefficients based on classifying the blind test set. The Matthews correlation coefficients, as with the full tripeptide motif features are low. For example the Laplace kernel with a cost of 1.0 trained on the full set of motifs has a Matthews correlation coefficient of 0.183, in comparison to 0.209 when trained on the reduced set.

Figure 2.18 shows the performance of an SVM classifier trained using the reduced tripeptide motifs as indicated by the solid black line. However, as indicated by the dashed lines, there is no difference from some SVMs that are trained using randomly generated data.

Figure 2.18: ROC plot for the final SVM trained with the Jpred Motif 27 feature set. The graph is plotted with 10 SVMs trained with random data (dashed line) to highlight the variability in the predictive capability when the SVM is trained with random data.



Tripeptide Sequence Motifs

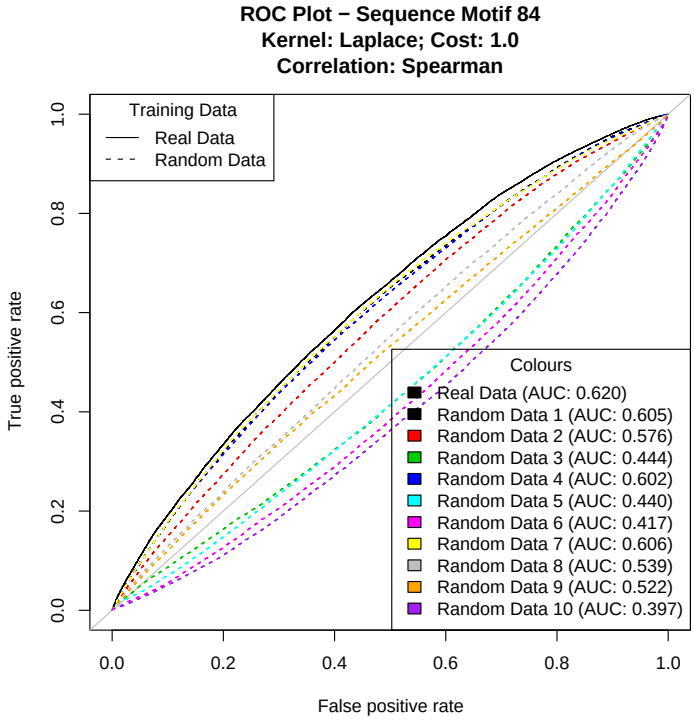
The top 20 results, ordered by the Matthews Correlation Coefficient are shown in Table 2.9 for the full motif set and Table 2.10 for the reduced motif set. The selected kernel for the classification of protein pairs based on sequence motifs was the Laplace kernel as it obtains the highest Matthews correlation. Both scaling and standardisation normalisation methods generated similar results (data not shown). The selected parameters are the Spearmans correlation, cost function of 1.0 and data normalisation via scaling. The ideal motif representation is the reduced 84 motif set as the Matthews correlations are over 3 times greater than the equivalent kernel trained based on the full 343 sequence motifs feature set.

When comparing an SVM trained with real information versus an SVM trained with randomly select data, the expectation would be for the SVM trained with real values to out perform the SVM trained with random data as assessed when tested with the same real examples. As shown in Figure 2.19 when comparing the performance of the selected SVM model with the random data, training with random data indicates that there is no significant difference between the SVM trained with real data and those trained with random data.

Proportions

The top 20 SVM models, ordered by the Matthews Correlation Coefficient are shown in Table 2.11. The Laplace kernel was selected for the classification of protein pairs based on sequence proportions. Both normalisation methods considered returned similar results (data not shown), therefore scaling was selected for further analysis. The final selected parameters were Spearmans correlation, cost function of 1.0 and

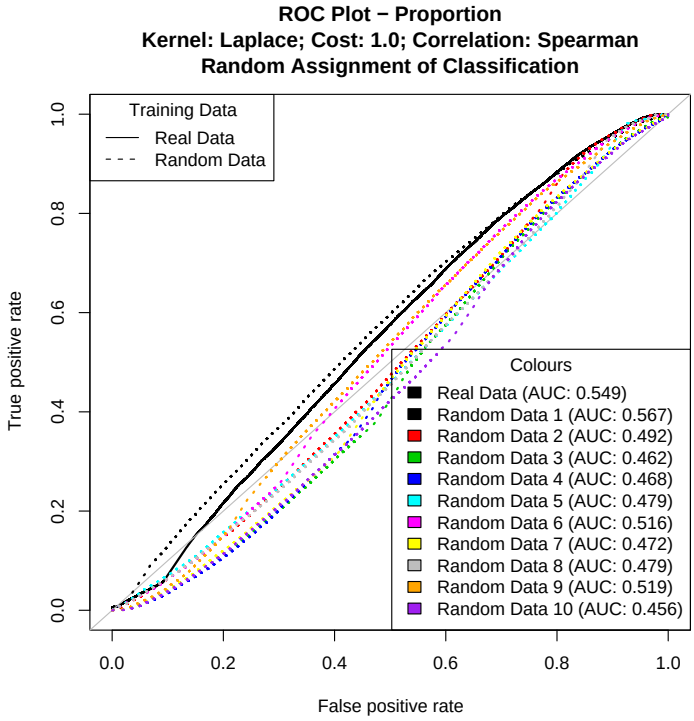
Figure 2.19: ROC plot for the final SVM trained with the Sequence Motif 84 feature set. The graph is plotted with 10 SVMs trained with random data (dashed line) to highlight the variability in the predictive capability when the SVM is trained with random data.



data scaled.

As with the SVMs trained with Jpred tripeptide motifs, both full and reduced feature sets, the Proportions of amino acids and structural features is slightly more effective at predicting protein-protein interactions than the SVMs trained with sequence motifs. One explanation for this is that more support vectors were required by the SVMs trained with sequence motif features over those trained with the proportion features. For example the SVM kernel Laplace and cost of 1.0 trained with 84 sequence features required 1861 support vectors in comparison to the same SVM parameters trained with sequence proportion features which required only 1670 support vectors. As shown in Figure 2.20, shows that most of the time SVMs trained with real data out perform SVM trained with a randomly generated training set.

Figure 2.20: ROC plot for the final SVM trained with the Proportion feature set. The graph is plotted with 10 SVMs trained with random data (dashed line) to highlight the variability in the predictive capability when the SVM is trained with random data.



Combined SVM Sequence Module

A combined prediction model can be generated based on the three SVMs. Table 2.12 describes the parameters selected to train each of the SVMs. The final predictor consists of 3 SVMs, the predictions are combined in a three dimensional binary array where each SVM predicts whether the two proteins interact or not (1 or -1 respectively). From the array it is then possible to calculate a final likelihood ratio to be incorporated into the PIPs framework.

The parameters shown in Table 2.12 for the Jpred Motif feature set show that the full Jpred motif feature set was selected rather than the reduced set, the reason for the selection is highlighted in Tables 2.13 and 2.14. Table 7a shows the calculated

Table 2.12: Parameters for the final 2x2x2 predictor.

Parameters	Reduced Sequence Motif (84)	Proportions	Jpred Motif (27)
Kernel	Laplace	Laplace	Laplace
Kernel Type	C-svc	C-svc	C-svc
Cost	1	1	1
Training Set (Pos:Neg)	1000:1000	1000:1000	1000:1000
Cross Fold Validation	5	5	5
Normalisation	Scaling	Scaling	Scaling
Correlation	Spearman	Spearman	Spearman

Table 2.13: Classifications of proteins based on SVMs trained with the Jpred Motif (10), Proportions and the Sequence Motif (84) for 33094 positive protein pairs and 33094 negative protein pairs.

Jpred Motif - 10	Proportion	Sequence Motif - 84	Positive Dataset	Negative Dataset	Likelihood Ratio
1	1	1	6737	2444	2.76
1	1	-1	3521	2075	1.70
1	-1	1	1327	720	1.84
-1	1	1	4669	3227	1.45
-1	-1	1	1972	2817	0.70
-1	1	-1	2425	2471	0.98
1	-1	-1	6294	6047	1.04
-1	-1	-1	6149	13293	0.46

likelihood ratios when combining the reduced Jpred motif feature set with the other two predictors. The maximum possible likelihood ratio (2.76) is less than when the SVM is trained with the full motif feature set in conjunction with the proportion and sequence motif feature sets.

To determine the effectiveness of the 3 SVMs, equivalent SVMs were generated and trained with random data. The ROC curves in Figures 2.18, 2.19 and 2.20 show that SVMs trained with known information result are above the diagoonal. Both SVMs trained with Jpred feature sets and Proportion feature sets have similar AUC values to equivalent SVMs trained with random data with a difference of less than 0.053. This indicates that the sequence module is less effective than the other two

Table 2.14: Classifications of proteins based on SVMs trained with the Jpred Motif (27), Proportions and the Sequence Motif (84) for 33094 positive protein pairs and 33094 negative protein pairs.

Jpred Motif - 27	Proportion	Sequence Motif - 84	Positive Dataset	Negative Dataset	Likelihood Ratio
1	1	1	5536	1232	4.49
1	1	-1	3981	1277	3.12
1	-1	1	1526	704	2.17
-1	1	1	4395	3520	1.25
-1	-1	1	1650	2779	0.59
-1	1	-1	2953	3323	0.89
1	-1	-1	6923	5661	1.22
-1	-1	-1	6130	14598	0.42

SVMs trained with real data.

Figures 2.18, 2.19 and 2.20 identify the range of SVMs that can be generated when an SVM is trained with random data. For the SVM trained with Jpred motif feature sets, the predictor remains above the diagonal of the ROC plot. However when the same SVM is trained with random data, the SVM is able to identify some interactions, but it is limited as shown by the plateau in the graph and then going below the diagonal. This work was taken no further due to the low predictive capability by each of the SVM modules and based on the low likelihood ratios of a combined module.

2.4 Conclusion

2.4.1 Combined Module

- It is possible to use the Gene Ontology to infer protein-protein interaction, although it does not provide enough evidence on its own to make a conclusive prediction.

- The integration of the Biological Process branch of the Gene Ontology in place

of co-localisation data showed improved performance in the Combined module.

- The PIPs 2 framework now includes the Gene Ontology as part of the Combined module in place of co-localisation

2.4.2 Clustering Module

- It is possible to predict protein-protein interactions by clustering a predicted interactome. This is competitive with the Transitive module.

- There is statistically not enough difference between the Clustering and the Transitive modules to include both within the PIPs framework and generate a single output. Both modules will be implemented within the PIPs framework, but there will be two output figures, one that has been calculated with the Transitive module and one with the MCL clustering module. The reason for calculating both scores is that even though the modules are not different enough to both be included together, there is still a large difference within the predictions that are made.

2.4.3 Expression Module

- Figure 2.15 shows the results for all of the human gene expression datasets considered. Over the range of 100 false positive results, the highest true positive rate is obtained when using the data derived from A-AFFY-44 chip with the expression set E-GEOD-7307 using Pearson as the measure of correlation, this has an ROC1000 AUC value of 256492.4, in comparison to E-TABM-145 (Spearmans) with an ROC1000 partial AUC score of 253606. Even though

the AUC scores are very similar the selected gene expression set and correlation measure is E-GEOD-7307 and Pearsons correlation. E-GEOD-7307 was selected over E-TABM-145 as it uses the A-AFFY-44 gene chip which has a larger coverage of the proteome.

- Even though in previous studies (Rhodes et al., 2005) expression correlation has been found to be highly predictive these results have been hard to repeat both in this and previous studies (Scott and Barton, 2007). That said, even though this module is not capable of making predictions on its own, it does provide evidence and a likelihood of interaction based on that evidence for a protein pair to interact and thus could be strong enough with other modules to predict whether two proteins interact or not.

- For future work, it might be worth considering different methods of biclustering gene expression data (Cheng and Church, 2000; Tanay et al., 2002; Gu and Liu, 2008), to identify sets of genes having similar expression patterns over a subset of conditions.

2.4.4 Sequencing Module

- Individually the SVM predictors based on Jpred and sequence tripeptide motifs and proportions of sequence features are weak, but combined, they perform slightly better at discriminating potential positive and negative interactions.

- The predictions from the Sequence module would never be strong due to the lack of accuracy and very low likelihood ratios, but what they would provide is greater coverage of the proteome.

- Due to the low likelihood ratios and the lack of accuracy implementing sequence information in this way, the Sequence module was not included within the PIPs 2 framework.

2.4.5 Updated Modules

The modifications to the Orthology and Transitive modules allow them to be run independently of the other modules for training and testing, although the Transitive module relies on the predictions of the Expression, Orthology and Combined modules, training and testing can be run without the requirement for retraining all the other modules. Code improvements to the Orthology module mean that the module is trained, tested and can make predictions quicker than PIPs 1. With the Transitive module now able to consider scores for all protein pairs when making final predictions, a larger space of protein-protein interactions can be considered.

Chapter 3

PIPs 2 Framework

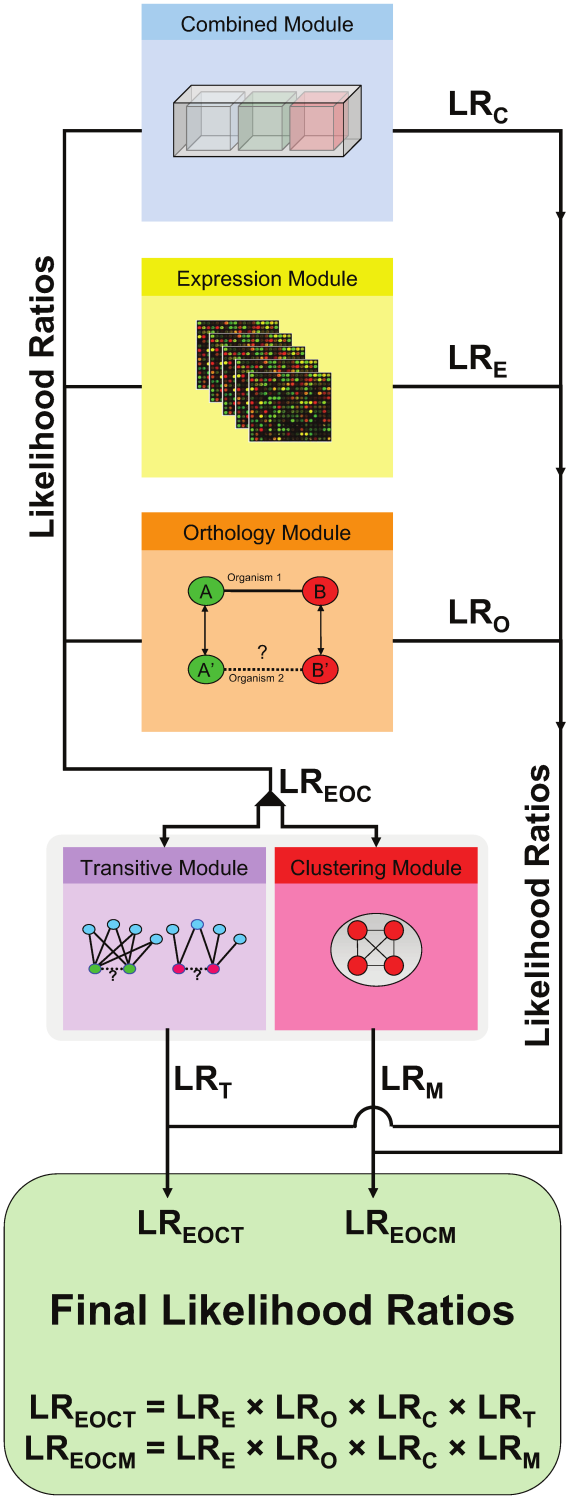
Preface

This Chapter describes the training and testing of the PIPs 2 predictor based on the developments and new modules described in Chapter 2.

3.1 Introduction

Chapter 2 investigated the development of new modules and improvements to existing modules within the PIPs framework. In Chapter 2 each of the modules was considered separately and independently of the others. Within a naïve Bayesian predictor it is the integration of independent modules simultaneously that increases the eventual predictive power and coverage. Figure 3.1 shows the structure of the PIPs 2 predictor based on the development of each module from Chapter 2. Figure 3.1 illustrates that the Expression (E), Orthology (O) and Combined (C) modules can be trained separately from the other modules within the PIPs framework. The Clustering (M) and Transitive (T) modules require the product of the likelihood ratios assigned by the E, O and C modules (EOC) to be trained, but can be trained

Figure 3.1: PIPs Framework Version 2. Each module is indicated by a coloured box (Blue, yellow, orange or purple). The arrows indicate how the likelihood ratios calculated by each module are combined. The final likelihood ratio for each protein pair is the product of the likelihood ratios calculated by each module. The Transitive and Clustering module use the product of the likelihood ratios from the Combined, Expression and Orthology modules for each protein pair to generate the local network of interactions.



independently of each other.

This chapter first describes the training and test sets for the full predictor as well as the calculation of the prior odds ratio. It then describes how the modules are combined, this Chapter also investigates the use of Support Vector Machines (SVMs) as an alternative to the naïve Bayes classifier to classify whether two proteins interact given precalculated likelihood ratios from each module. If a protein pair that is known to interact only has a single source of information, the naïve Bayesian approach would be less likely to identify that protein pair as interacting, unlike an SVM based method. Support Vector Machines (SVMs) can be used to combine the predictions made by each of the module in a different way than that of the naïve Bayesian method by identifying multi-dimensional correlations based on the predictions by each of the modules. Section 3.3 deals with measuring the accuracy of the predictor, the predictions that are made, the use of SVMs for classification and then the limitations of the predictor.

3.2 Methods and Data Sources

3.2.1 Training and Testing

The HPRD (Keshava Prasad et al., 2009) was chosen as the source of the positive examples because this database is derived from manual curation of the literature to identify protein-protein interactions from both high and low throughput methods, plus the HPRD contains a large number of interactions (see Section 1.4). The training set consists of 33,309 binary protein-protein interactions from the HPRD.

Given that the complete interactome is unknown, it is not possible to guarantee

that any selected pair of proteins do not interact. For the generation of a negative dataset for training it is necessary to select pairs of proteins that are unlikely to interact, or at least reduce the number of known interacting pairs of proteins. Several methods of pair selection are possible:

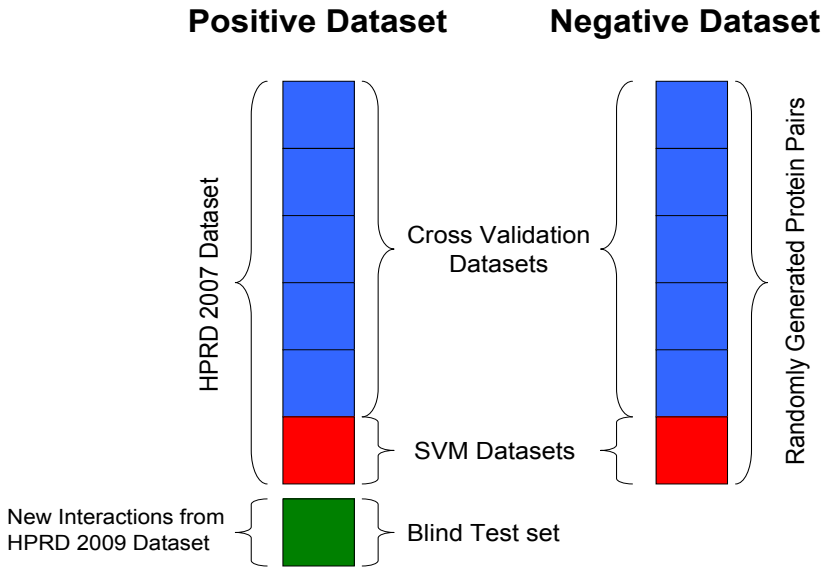
- Proteins in separate compartments
- Random selection and filtering
- Predictive methods (Shoyaib et al., 2009)

Each method is discussed further in Section 1.5.2, however random selection of protein pairs and then filtering out known or predicted interactions between protein pairs was selected to generate the negative dataset. Random selection of protein pairs was chosen to avoid introducing a bias in the negative training set. If protein pairs had been selected based on occurring in different compartments within the cell it could cause the predictor to infer that if a protein pair are co-localised then they are going to interact even though not all protein pairs interact even if they are present in the same compartment. The randomly selected pairs were filtered to remove proteins pairs that are know or predicted to interact based on interactions within the HPRD (Keshava Prasad et al., 2009), BioGRID (Stark et al., 2006), DIP (Salwinski et al., 2004), IntAct (Aranda et al., 2009) or OPHID (Brown and Jurisica, 2005) as done previously (Scott and Barton, 2007).

The predictor was trained on 33,309 binary protein-protein interactions, which is an increase of 6413 interactions on the previous PIPs predictor, plus there has also been an increase in the coverage of the proteome from 22,889 proteins to 25,674.

The number of negatives that was used during training was 100 times larger than the positive set to represent the low probability of interaction for randomly chosen proteins. A training ratio of 1:1000, based on the prior odds ratio (Section 3.2.2) would have been better, but this would have required more memory than was available. The complete training/testing dataset was divided into 6 sets, as shown in Figure 3.2; sets 1 to 5 (Figure 3.2, blue) totalling 27,757 annotated interactions were used for 5 fold cross validation; the 6th set consisting of 5552 interactions, was used for the training and testing of the application of Support Vector Machines to classify protein-protein interactions. All 6 sets were used for training the final predictor. In comparison to version 1 of the PIPs predictor, the cross fold validation sets are pre-determined prior training and testing, unlike before where they were randomly selected for each training run. This allows for the results to be repeatable and for easy comparison for further development of new modules.

Figure 3.2: Division of datasets for training of the PIPs Predictor. The datasets were divided into 6 sections; the blue section was used for 5 fold cross validation of the PIPs predictor modules; the red section was used for training and testing the SVMs.



The final blind test was made of the new interactions present in the HPRD 2009 dataset and were not considered during training and testing of the PIPs predictor. In addition to the 33,309 interactions that were used during training and testing, a further 2678 interactions had been derived from the subset of the HPRD 2009 dataset that were not present in the HPRD 2007 dataset.

Pre-generating the training and test sets allows for easier construction and testing of modules that utilise the pre-computed likelihood ratios from the co-expression, orthology and combined modules, such as the Transitive and Clustering modules. Pre-calculating the modules means that the Transitive and Clustering modules can be developed without requiring the Expression, Orthology and Combined modules to be computed on the fly which is computationally expensive and increases the time it takes to evaluate changes that are made within the Transitive and Clustering modules. The end result is to allow for a more rapid development of new modules within the PIPs framework.

3.2.2 Setting the Prior Odds Ratio, O_{prior}

The prior odds ratio (O_{prior}) is the estimate of how many times more likely a given event is to occur than to not occur by chance. Therefore, setting the O_{prior} determines how high the likelihood ratio between a protein pair has to be before that interaction is more likely to occur than to not occur. Table 3.1 shows the number of proteins and the number of interactions present in the 2007 and 2009 editions of the HPRD database. Table 3.1 also shows the number of interactions between the set of proteins in the 2007 HPRD dataset and the number of interactions between the same set of proteins in the HPRD 2009 dataset. These values show that the

Table 3.1: Calculation of O_{prior}. The number of protein-protein interactions in the 2007 and 2009 releases of the HPRD along with the increase in the number of interactions between the proteins present in both releases. $2007 \cap 2009$ is the subset of protein present in both datasets and the number of interaction between proteins within the subset based on the 2009 dataset

HPRD	Proteins	Interactions	O_{prior}
2007	8968	33309	$\frac{1}{1206}$
2009	9177	35025	$\frac{1}{1201}$
$2007 \cap 2009$	8968	33490	$\frac{1}{1200}$

estimates of the O_{prior} are tending to get lower as the coverage of the interactome increases, but are around $\frac{1}{1200}$. Due to the network being incomplete and there not being a full coverage of proteins within the set a conservative estimate of $O_{prior} = \frac{1}{1000}$ has been used in this work. An O_{prior} of $\frac{1}{1000}$ is more stringent than the previous predictor ($\frac{1}{400}$ (Scott and Barton, 2007)) and as stringent as what was proposed as a potential O_{prior} within the paper ($\frac{1}{1053}$ (Scott and Barton, 2007)).

3.2.3 Database

The annotations used by each of the modules are described below.

Combined Module

The Combined modules includes the similarity of Gene ontology terms, co-occurrence of post-translational modifications and protein domains (see Section 1.8.1 for a description of the Combined module in PIPs version 1 and Section 2.2.1 for the introduction of GO terms). The Gene Ontology terms were downloaded from the GOA website (Barrell et al., 2009). 17266 human proteins had Biological Process GO term annotations. Domain annotations were derived from InterPro (Hunter et al.,

2009) and Pfam domains that are also known to interact were also considered (Jefferson et al., 2007) as per PIPs 1. There are 6254 distinct proteins that have post translational modification annotations.

Expression Module

The expression module uses the dataset E-GEOD-7307 downloaded from the ArrayExpress database (Parkinson et al., 2009).

Orthology Module

Orthologous proteins were downloaded from the InParanoid database (Berglund et al., 2008). There are 32,981 distinct orthologs between human and yeast, fly or worm, covering 10,848 distinct human proteins. The interactions that were present in yeast, fly and worm were downloaded from DIP (Salwinski et al., 2004) and IntAct (Kerrien et al., 2007a; Aranda et al., 2009).

Transitive and Clustering Module

The Transitive and the Clustering modules use protein pairs whose product of likelihood ratios calculated by the Expression, Orthology and Combined modules are greater than or equal to a set threshold. The threshold for the Transitive module is set at a likelihood ratio of ≥ 10 as has been done previously (Scott and Barton, 2007). The threshold for the Clustering module was set lower than the Transitive module at a likelihood ratio of ≥ 5, as described in Section 2.2.2.

3.2.4 Naïve Bayesian Classification

To combine the modules in a naïve Bayesian manner the final likelihood ratio for a protein pair is the product of the likelihood ratios calculated by each of the modules. The protein pair are therefore predicted to be more likely to interact if the product of the likelihood ratio and the O_{prior} is ≥ 1.0. Further details are provided in Chapter 1.6.1.

3.2.5 SVM Classification

To investigate the use of SVMs in the prediction or protein-protein interactions likelihood ratios were calculated for protein pairs in dataset 6 (see Figure 3.2). The predictions for Dataset 6 where made by training PIPs 2 on datasets 1 to 5 and then calculating predictions for the protein pairs in dataset 6. Dataset 6 consisted of 5552 known protein-protein interactions as the positive set and 588111 non-interacting protein pairs as the negative set.

For assessing set size for training an SVM, increasing numbers of positive and negative examples were used for training. For each SVM training set of different sizes, dataset 6 was sampled to extract examples of positives and negatives that were split into 5 non-overlapping groups which were used for 5 fold cross validation. However the test set was a randomly selected sample of 1000 positive and 1000 negative examples from Dataset 6 that were not used for training. This allowed for an error to be assigned to the trained SVMs for all the different training set sizes considered.

Assessment of the effect on the ratio of positive to negative examples used in SVM training was performed by generating two negative datasets, one of equal size

to the number of positive set and one that was 10 times larger. Each set was divided into 5 and used for five fold cross validation. The 1:1 positive to negative training sets were also used for determining the effect of different methods of normalisation.

SVM Dataset Normalisation

Several data normalisation methods were compared to identify the most predictive (see Table 3.2). Equation 3.2.1 is the z-score.

Table 3.2: Methods of normalisation used.

Method ID	Desciption
A	No scaling, raw likelihood ratio values.
B	Equation 3.2.3 (Log)
C	Equation 3.2.2 (Linear)
D	Equation 3.2.1 (Standard Score)

$$v_i' = \frac{(v_i - \bar{V})}{\sigma_V}$$

Equation 3.2.1: Standard score (z-score)

where v_i is an element of V, the vector of likelihood ratios calculated for Dataset 6 by a given module for each protein pair and v_i' is the normalised value of v_i. $\bar{V}$ is the mean and σ_V is the standard deviation of all values within the set V. This provides a way of scaling the data so that it is within a reasonable range. Values such as this are then theoretically ideal for passing into an SVM.

Equation 3.2.2 involves normalising the data so that the values are linearly scaled between 0 and 1.

where v_i is an element of V, the set of all likelihood ratio values. This ensures that all of the values are scaled to between 0 and 1.

$$v_i' = \frac{(v_i - \min(V))}{(\max(V) - \min(V))}$$

Equation 3.2.2: Linear Scaling

Table 3.3: settings to select method of normalisation.

Kernel	Radial Base Function
Classification Method	C-svc
Number of k-fold Validation Rounds	5
Cost Function	1
Total Number of Training Examples	8882 (4441 positive, 4441 negative)

Equation 3.2.3 involves taking the log of the final likelihood ratio.

$$v_i' = \log_{10}(v_i)$$

Equation 3.2.3: Linear Scaling

SVMs were also trained on data that had not been normalised to identify if normalisation greatly improved the predictive capability of the models.

Support Vector Machine

Classification was done using the kernlab package within the R programming framework. The settings that were used for the support vector are shown in Table 3.3.

3.3 Results

3.3.1 Combining Modules

Each of the modules was assessed individually to determine their predictive capability. A naïve Bayesian approach requires all modules to be independent. The independence of the modules was assessed by pairwise Pearsons correlation of each

pair of modules and the results are shown in Table 3.4. Due to limitations in the size of the matrix that can be held within the R environment only predictions where likelihood ratios were greater than 1.0 (45.34M protein pairs) were considered.

Table 3.4: Correlation between final predictions made by each module.

Module	Expression	Orthology	Combined	Transitive	Clustering
Expression	-	-0.0047	-0.0370	-0.0011	0.0058
Orthology	-	-	0.0201	0.0985	0.0620
Combined	-	-	-	0.1162	0.0777
Transitive	-	-	-	-	0.2494
Clustering	-	-	-	-	-

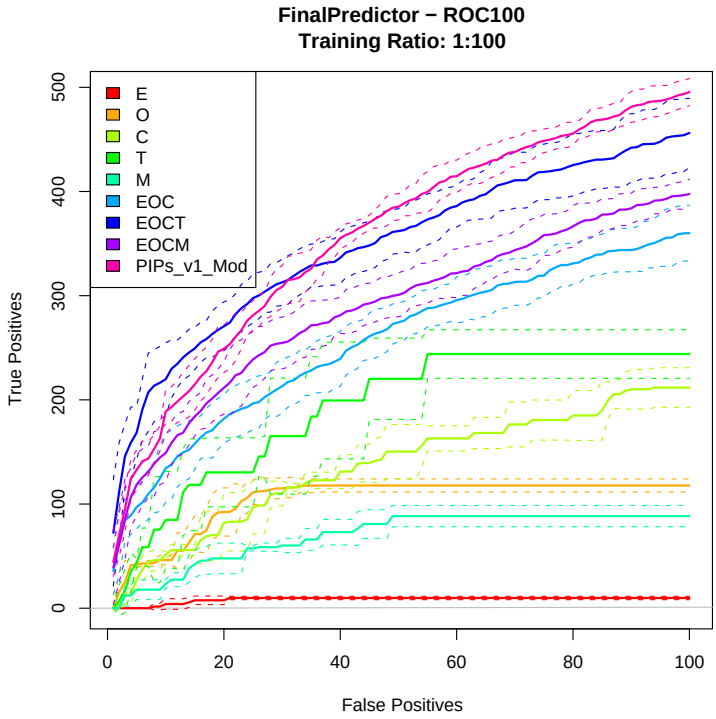
Table 3.4 shows that there is a correlation of 0.25 between the Transitive and Clustering modules indicating there is an overlap in the predictions made. To avoid over estimation of the likelihood of interactions between some protein pairs, these modules were never considered simultaneously. All other module pairs gave Pearsons correlations between -0.01 and 0.12 and are therefore regarded as uncorrelated. Figure 3.1 shows the final structure of PIPs framework, with the final predictor calculating two predictions based of Expression, Orthology, Combined and either the Transitive or Clustering modules (EOCT or EOCM respectively).

3.3.2 Accuracy of PIPs 2

Figure 3.3 shows that all the modules have ROC100 curves that are greater than random. The maximal ROC100 values, based on fixed 5 fold cross validation, for EOC, EOCT and EOCM are 360, 460 and 400 respectively. This represents a slight reduction over PIPs version 1 predictor (500), but is not a significant drop as the values are within 1 standard deviation of each other. This slight reduction in the maximal ROC100 values is compensated by a large increase in the coverage of the

proteome and in the final number of predictions for the new EOCT and EOCM predictors.

Figure 3.3: ROC100 plots for the Final PIPs predictor along with the ROC100 plots for the individual modules based on five fold cross validation. E: Expression; O: Orthology; C: Combined; T: Transitive; M: Clustering. EOC is the combined predictive accuracy of the E, O and C modules, likewise EOCT and EOCM is the combined predictive accuracy of the EOC and T and M modules respectively. The pink line is the accuracy of the PIPs 1 predictor. The dotted lines are 1 standard deviation based on variance of true positive predictions made per false positive prediction during 5 fold cross validation. The grey line represents random selection



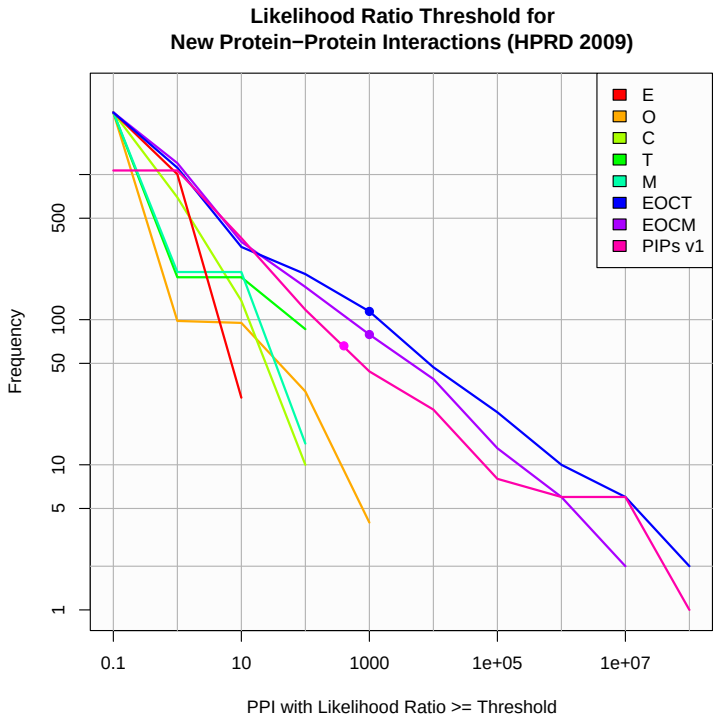
A blind set of protein-protein interactions was generated to analyse the predictive capability of the PIPs predictor, as illustrated in Figure 3.2. Figure 3.4 shows the number of predictions that are made by PIPs 2 EOCT and EOCM in comparison to the predictions that are made by PIPs 1 (EOCT) of interactions that are in the HPRD 2009 dataset but not in the HPRD 2007 dataset. Table 3.5 shows the number of predictions for each of the sets where the protein pair has a likelihood

Table 3.5: Number of predicted interactions within the Blind test set.

HPRD Dataset	2009
Total No. Interactions (Non Homodimers)	38213 (36100)
HPRD 2009 \ HPRD 2007 Non Homodimers	2678
PIPs 1	66
PIPs 2 Union	137
PIPs 2 EOCT	114
PIPs 2 EOCM	79

ratio greater than or equal to the threshold point (likelihood ratio 400 for PIPs 1 and likelihood ratio 1000 for PIPs 2). Even though the PIPs 2 predictors use a more stringent prior odds ratio of 1000 compared to the prior odds ratio for PIPs 1 of 400, the PIPs 2 predictors together predict over twice as many interactions as PIPs 1 in the blind test set.

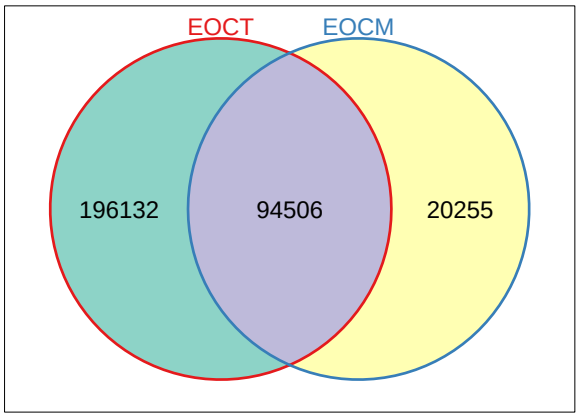
Figure 3.4: Blind set predictions of new protein-protein interactions present in the HPRD 2009 database, but were not present in the HPRD 2007 dataset, which were used for training and testing.



Predictions

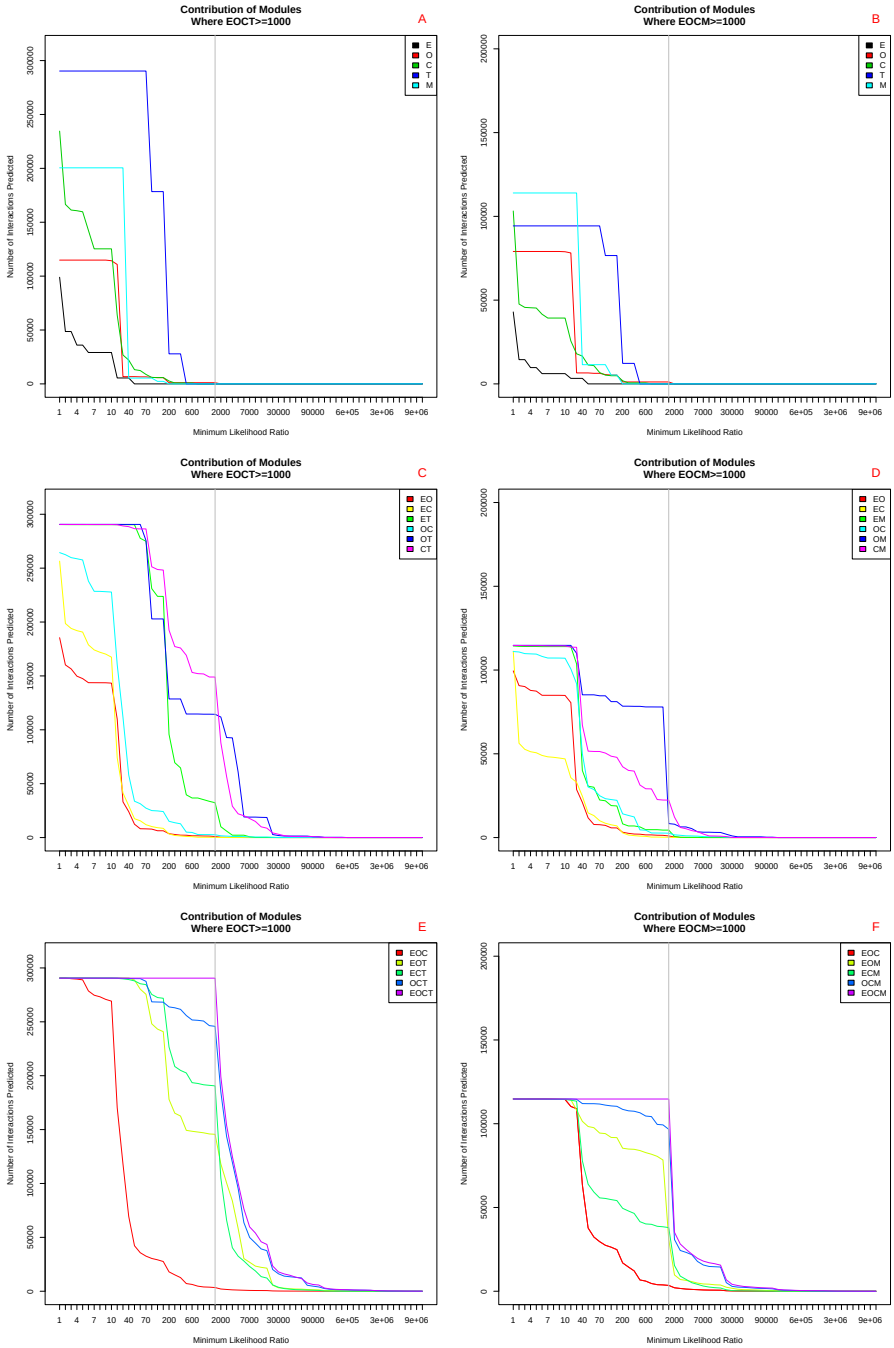
For PIPs 2, the EOCT and EOCM predictors both calculate a larger number of potential protein-protein interactions with a likelihood ratio 1000 (290,638 and 114,761 respectively) than EOC (3519). Figure 3.5 shows that the overlap of EOCT and EOCM is 94,506 predicted protein-protein interactions.

Figure 3.5: Venn Diagram of the number of interactions predicted by EOCT and EOCM as part of PIPs version 2 and the intersect of the two sets of predictions.



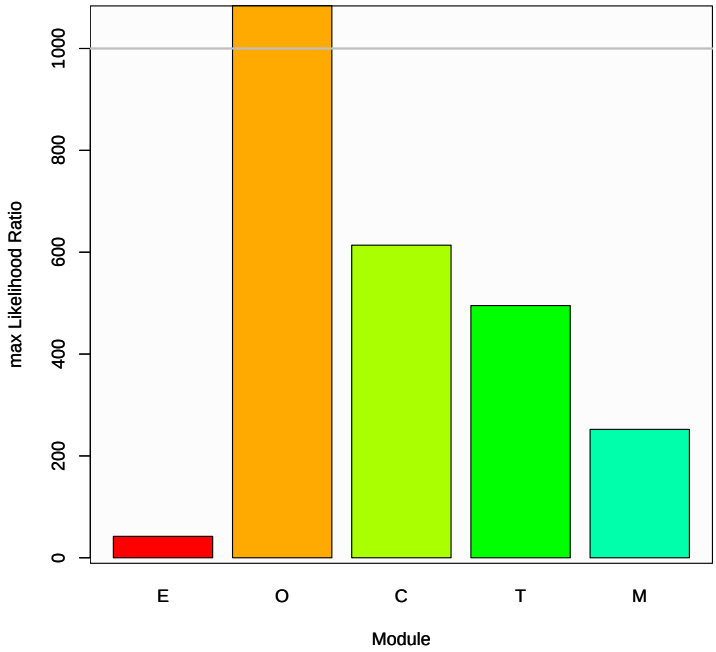
With a naïve Bayesian approach is it possible to identify the modules that had the largest influence on the final set of predictions. Figure 3.6 shows graphs for the breakdown of the contributions that are made by each of the modules for the EOCT and EOCM predictor where their interactions have likelihood ratios ≥ 1000. The Orthology module is the only module capable of assigning a likelihood ratio ≥ 1000 (see Figure 3.7) and is therefore capable of making a prediction on its own. This module will only function if the protein pair has at least two sources of evidence from different species with orthologous / paralogous interacting protein

Figure 3.6: Break down of the contributions made by each module to the final set of predicted protein-protein interactions with likelihood ratios ≥ 1000 by EOCT (A, C and E) and EOCM (B, D and F) where the modules are labelled as: E: Expression; O: Orthology; C: Combined; T: Transitive; M: Clustering. Panels A and B show the number of predictions for each module individually. Panels C and D show the number of predictions based on the combination of two modules. Panels E and F show the number of predictions made by three modules and with all four modules combined.



pairs. The Clustering (M) and Transitive (T) modules provide the largest coverage of the predicted interactome; this is because they consider the sum of all unique proteins that are analysed by the Expression, Orthology and Combined modules and therefore not constricted to a single source of evidence. For both the EOCT and EOCM predictors, the largest contributions to the final set of predictions are made by the Orthology, Combined and Transitive or Clustering modules. As with PIPs 1 (Scott and Barton, 2007), the Expression module has the lowest maximum likelihood ratio (42.0).

Figure 3.7: Maximum likelihood ratios that can be assigned by each of the modules. The grey line indicates the likelihood ratio required to predict that a protein pair is more likely to interact than to not interact.



PIPs 2 also shows a large increase in the number of predictions that were made. This is because of the restructuring of the predictor so that the Expression, Orthology and Combined modules were able to make a full set of predictions, which

could then be used to generate a full EOC network of protein-protein interactions that could be considered by the Clustering and Transitive modules. In PIPs 1 this was not possible as the module had to be trained in series which would result in a limited number of interactions that could be considered by a Network module.

3.3.3 Comparison of Predictions Made By PIPs 2

PIPs 2 vs PIPs 1

Based on the ROC100 plot in Figure 3.3, there is no significant difference between the PIPs 1 predictor and PIPs 2 EOCT predictor at a false positive count of 100, however, PIPs 2 EOCT outperforms the PIPs 1 predictor at lower false positives of 20 where the Likelihood ratio threshold for PIPs 2 $\geq$ 1000. For the first 20 false positives the PIPs 2 predictor is greater than 1 standard deviation from PIPs 1. There is a significant difference between EOCM and EOCT ($\geq$ 1 standard deviations in the predictive rate.). However, EOCM represents a new method and its predictions targets different protein pairs than the EOCT modules, making it complementary with a Pearson correlation of 0.25 and the overlap of the EOCT and EOCM predictions shown in Figure 3.5. This indicates that EOCT and EOCM are predicting in different areas allowing for an increase in the coverage of the predictions that can be made by the PIPs 2 predictor.

There is an overlap of 8669 predicted protein-protein interactions between PIPs1 and PIPs 2, which is an overlap of 23.05%. This small overlap between the two predictors may be due in part to changes in the experimental evidence that is now accessible to PIPs; there is a change of the gene expression data as part of the Expression module, different interaction databases are considered by the Orthology

module and the Combined module now considers the semantic similarity of GO terms rather than colocalisation. There have also been changes to the source databases for protein annotations.

PIPs 2 vs Other Databases

Table 3.6 highlights the overlap between PIPs 2 and other protein-protein interaction databases, both experimentally determined and predicted. Proportionately, PIPs 2 has the largest overlap with DIP (18%) and the lowest with IntAct (5%). Even when comparing to OPHID there is only an overlap of 9%. This is in agreement with what had been found previously and that there is still a low level of overlap between protein-protein interaction databases (Scott and Barton, 2007).

3.3.4 SVM Classification

Table 3.7 shows the results for training using different methods of normalisation, but none were found to be significantly different. Equation 3.2.3 was selected out of the methods considered for normalisation of the data. The difference between each normalisation method is not much, with differences of Matthews Correlations within 0.1 of each other. All of the methods required 5000 support vectors to generate a

Table 3.6: Overlap between PIPs 2 and various protein-protein interaction databases (known and predicted). The number of interactions is of non-self interacting interactions.

Database	No. Interactions	Intersection With			
		EOCT	EOCM	EOCT ∩ EOCM	EOCT ∪ EOCM
DIP	1215	174	146	94	226
HPRD	36100	3163	2111	1549	3725
IntAct	17456	850	570	442	978
OPHID	81259	6453	4075	2773	7755

viable model of the data.

Table 3.7: The results for the SVMs generated using the different normalisation methods.

Method ID	TP	FP	FN	TN	Number of Vectors	Cross Fold Error	Matthews Correlation Coefficient
A	726	182	384.4	928.4	5210.4	0.25	0.50
B	780.8	190.8	329.6	919.6	4964.8	0.23	0.54
C	826.4	292.2	284	818.2	5110.2	0.26	0.48
D	829	321.8	282	789.2	5218.6	0.25	0.46

Table 3.8 shows that viable SVM models are capable of being produced with this data independent of the number of examples that are available during training. Therefore if the data is available, then it can be used for training SVM's, but in circumstances when training examples are limited then SVMs are resistant to this lack of information.

Table 3.9 shows that altering the training ratio does reduce the Matthews correlation coefficient by 0.07 when going from 1:1 to 1:10. Hence the use of a balanced training dataset of positive to negative examples is preferable.

However, when comparing the classifications that are made by the naïve Bayesian classifier, there is a significant advantage to the use of SVMs for combining the modules rather than taking the product of the likelihood ratios calculated by each of the modules. The Matthews correlation when setting the $O_{prior} = 1000$ is 0.16 with a standard deviation of 0.005; this indicates that the naïve Bayesian method performs significantly better than random (Matthews correlation of 0). However, the Matthews correlation for SVMs using Log10 normalisation is 0.54 with a standard deviation of 0.02. When the trained SVM module is then used to classify the blind

Table 3.8: Effect on SVM model calculation dependent on the size of the training set. The total size is the sum of positive and negative examples (ratio 1:1)

Total Number of Examples	TP	FP	FN	TN	Number of Vectors	Cross Fold Error	Matthews Correlation Coefficient
200	738	215.2	262	784.8	1374.8	0.25	0.52
1000	733.4	190	266.6	810	1779.2	0.24	0.55
2000	695	166.8	305	833.2	2368.2	0.24	0.53
3000	700.2	170.2	299.8	829.8	2925.4	0.24	0.54
4000	714.8	196.2	285.2	803.8	3403.4	0.24	0.52
5000	725.4	202	274.6	798	3906.4	0.24	0.53
6000	741.2	178.8	258.8	821.2	4412.6	0.23	0.57
7000	719.6	199.4	280.4	800.6	5059.8	0.24	0.52

Table 3.9: The effect on SVM model generation due to altering the bias of positive to negative examples.

Train Ratio	TP	FP	FN	TN	Number of Vectors	Cross Fold Error	Matthews Correlation Coefficient
1:1	782.2	191.8	328.2	918.6	4959.8	0.23	0.54
1:10	352.4	102.2	758	11001.8	7599.6	0.07	0.47

test set it is capable of identifying 1875 protein-protein inteactions (standard deviation 19), which is much larger than that of the final PIPs 2 predictor using a naïve Bayesian method. The reason for this is that the naïve Bayesian method takes a simplistic approach to combine the predictions that are made by each of the modules assigning each an equal weight, therefore it is not able to identify subtleties within the predictions where the product of the likelihood ratios is above a fixed threshold. The result of the naïve Bayesian method is that if a protein pair has only a single source of evidence it might not classify a given protein pair as interacting. SVMs are capable of handling subtle patterns that could indicate whether two proteins are likely to interact or not, even at much lower likelihood ratios, in ways that the naïve Bayesian method is not able to with a fixed threshold (O_{prior}).

3.3.5 Limitations

Due to the statistics involved in calculating likelihood ratios, it is not possible to identify a likelihood ratio for homo-dimers. The PIPs framework would consider a pair of proteins that have identical information to be more likely to interact than to not, irrespective of whether the protein is able to interact with itself or not. Therefore the PIPs predictor is only able to consider non-self interactions. This is a limitation for the predictor as many biological units require self interaction to become functional.

Another limitation with the PIPs predictor is that it is only able to consider proteins that have experimental or annotative evidence. This means that not all proteins are treated equally, as some proteins will have several isoforms, but only one of the isoforms will have been studied. The limitation is a result of the coverage of the

proteome by experimental evidence. However, with high throughput experiments becoming faster and cheaper, the availability of data for the whole proteome should become more accurate and more accessible. The increase in information about the proteome will allow methods, such as the PIPs framework, to interrogate more protein pairs to predict whether they are likely to interact. Already the number of predictions is over 300,000; this is within the predicted size of the interactome which ranges between 130,000 and 650,000 (Hart et al., 2006; Stumpf et al., 2008; Venkatesan et al., 2009), but those estimates did not take into account the total number of protein isoforms for each gene, therefore the actual number of unique interactions could be much larger. The Sequence module was developed to target this limitation in the predictor (see Section 2.2.4), however it was not found to be predictive.

3.4 Discussion and Conclusions

The new PIPs predictor, providing two final predictions (EOCT and EOCM) is more capable of identifying new interactions than the previous predictor (Figure 3.4). Along with the alterations to the way that the predictor functions, each of the primary evidence module (Expression, Orthology and Combined) can make predictions independently allowing the Transitive and Clustering modules to analyse the full set of EOC predictions thus making the predictor much more effective. As well as the modifications to the modules and the functioning of the predictor as a whole there is also the increased coverage of the proteome, from 18k to 25k proteins.

SVM classifiers are capable of identifying positive and negative protein-protein

interactions given PIPs module likelihood ratios. Further development of this module has not been taken forward due to time constraints, but it is possible to implement this method to classify whether two proteins are likely to interact or not independent of whether they have a posterior odds ratio of ≥ 1.0.

- PIPs 2 makes two final predictions (EOCT and EOCM), which is more capable of identifying new interactions than PIPs 1 (see Figure 3.4).

- Alterations to how the modules are trained and the way that they make predictions means that the Expression, Orthology and Combined modules can be run independently of each other, therefore allowing them to be run in parallel. All of the predictions made by the Expression, Orthology and Combined modules can then be considered by the Transitive and Clustering modules, unlike in PIPs 1, making them more effective.

- There is an increase in the coverage of the proteome by the PIPs 2 predictor from 18k to 25k proteins.

- It is possible to use SVMs to combine the likelihood ratios calculated by each module to make the final prediction of interaction between protein pairs.

- Due to the low overlap between the PIPs 1 and PIPs 2 predictors, training PIPs 2 on the datasets used for PIPs 1 would provide help identify improvements in the predictor.

Chapter 4

Analysis of PIPs 2 Predictions

Preface

Based on the PIPs 2 predictor described in Chapter 3, this Chapter goes on to investigate the full set of predictions that have been calculated.

4.1 Introduction

This chapter covers the analysis of the protein-protein interaction predictions that have been made by the PIPs 2 predictor (see Chapter 3 for further details). Section 4.2 assesses the accuracy of the PIPs predictions with protein pairs that have been determined not to interact. Sections 4.3.1 and 4.5 analyses the enrichment of interaction between sets of protein that are part of defined groups, such as co-localisation, co-complexed or are part of a similar functional pathway. Section 4.4 highlights biologically significant interactions that have been predicted by the PIPs 2 predictor and Section 4.6 suggests how to experimentally validate the predictions that have been made.

The PIPs 2 predictor was used to calculate the likelihood ratios of over 300M

protein pairs from the IPI where both proteins had at least 1 form of evidence, of which 310,894 interactions have a calculated likelihood ratio ≥ 1000 by either EOCT or EOCM, this is known as the LR1000u set. 94,507 protein pairs have both EOCT and EOCM likelihood ratios ≥ 1000, this is known as the LR1000i set. The two sets cover 12,633 and 5553 distinct proteins respectively indicating that there are many proteins that lack sufficient evidence to be predicted to interact with a final likelihood ratio of ≥ 1000 with another protein.

4.2 Comparison of PIPs 2 Predictions to Known Negative Interactions

One of the many difficulties that arise from the prediction of protein-protein interactions is the selection of a negative training set. It is rare for publications to highlight negative protein-protein interactions, it is then even rarer for those annotations to make it into a database that is publicly accessible and easily machine readable.

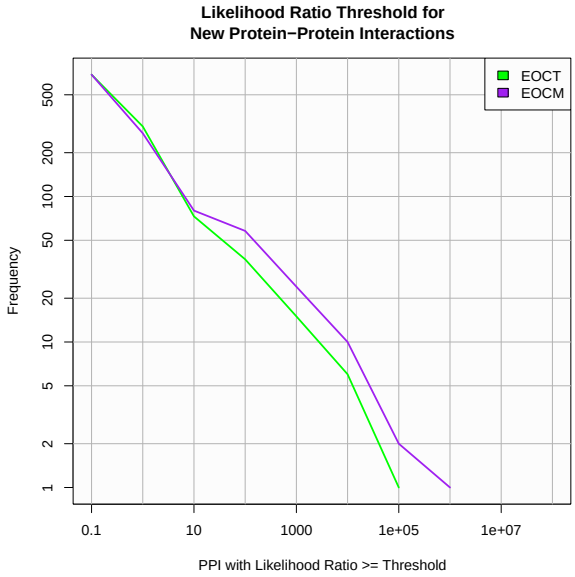
Recently, protein-protein interaction databases have started to recognise the issue of identifying protein pairs that do not interact as being as important as identifying those proteins that do interact and have started to publish databases of protein pairs that do not occur. IntAct (Kerrien et al., 2007a; Aranda et al., 2009) has introduced a new tag to identify such protein pairs where researchers have found evidence to show that two proteins, in a defined set of experimental conditions do not interact. The Negatome database (Smialowski et al., 2009) also reports proteins that do not interact. The non-interacting proteins often come from the literature where the protein pairs have been selected as part of controls for experiments or

they are proteins that have been targeted to identify interactions. The Negatome also selects proteins that are part of X-ray crystallographic complexes in which two proteins are found to be separated by a third protein. The two separate proteins are annotated as not interacting.

The Negatome contains two manually curated sets of non-interacting protein pairs, one is filtered against IntAct (manual-stringent) and the second is not (manual). The manual stringent set comes from protein pairs that have been annotated in the literature to identify the lack of an interaction between a pair of proteins; this set has then been filtered against the IntAct database to remove subsequently identified interactions. To compare this set to the PIPs predictions, the selected Negatome set was then filtered to remove protein pairs that are listed as part of the same complex by the HPRD (Mishra et al., 2006; Keshava Prasad et al., 2009). Filtering the Negatome set of non-interactions is done because it is harder to identify the true non-interactions within co-complexed sets of proteins as they will often have enriched likelihood ratios due to similar annotations and complementary gene expression patterns (see Section 4.3.1). After filtering this leaves 688 protein pairs that are annotated as not interacting out of 1291 negative interactions.

Figure 4.1 shows how many of the 688 negative interactions have assigned likelihood ratios greater than or equal to a given threshold. The graph highlights that at a likelihood ratio threshold of 1000, there are 24 EOCT and 15 EOCM negative interactions. There are 25 negative interactions where either EOCM or EOCT had a score greater than or equal to 1000, these are listed in Table 4.1. Many of these protein pairs are involved in similar processes which makes the task of identifying the two proteins as not interacting more difficult. In the Peroxisomal membrane

Figure 4.1: Cumulative frequency graph of Negatome interactions and their corresponding likelihood ratios as calculated by PIPs.



complex (PEX) there are many proteins within the complex, but due to the structural configuration of the complex, many of the proteins do not interact, such as PEX11A and PEX11B. The interaction between PRPF3 and PRPF4 has also not been confirmed to occur (Liu et al., 2006), the two proteins are part of the same complex as part of the spliceosome, but they have not been experimentally validated to interact.

Table 4.1: Negative protein-protein interactions present within the Negatome, but predicted to interact by PIPs.

Protein 1	Description	Protein 2	Description	EOCT	EOCM
CDC45L	CDC45L CDC45-related protein	MCM2	MCM2 DNA replication licensing factor MCM2	1.22E+06	197808
PEX12	PEX12 Peroxisome assembly protein 12	PEX13	PEX13 Peroxisomal membrane protein PEX13	25561.4	47636
PEX12	PEX12 Peroxisome assembly protein 12	PEX11A	PEX11A Peroxisomal membrane protein 11A	14762.6	27511.5
PEX11A	PEX11A Peroxisomal membrane protein 11A	PEX13	PEX13 Peroxisomal membrane protein PEX13	14762.6	27511.5
PRPF3	PRPF3 Isoform 1 of U4/U6 small nuclear ribonucleoprotein Prp3	PRPF6	PRPF6 Pre-mRNA-processing factor 6	118026	25539.8
BCL2L2	BCL2L2 Apoptosis regulator Bcl-W	MCL1	MCL1 Isoform 1 of Induced myeloid leukemia cell differentiation protein Mcl-1	12428.1	12529.4
PRPF3	PRPF3 Isoform 1 of U4/U6 small nuclear ribonucleoprotein Prp3	TXNL4A	TXNL4A Thioredoxin-like protein 4A	45376.3	9819.1
SART1	SART1 U4/U6.U5 tri-snRNP-associated protein 1	PRPF4	PRPF4 Isoform 1 of U4/U6 small nuclear ribonucleoprotein Prp4	42915.2	9286.5
PRPF3	PRPF3 Isoform 1 of U4/U6 small nuclear ribonucleoprotein Prp3	PRPF4	PRPF4 Isoform 1 of U4/U6 small nuclear ribonucleoprotein Prp4	29466.9	6376.4

Continued on Next Page...

Table 4.1 – Continued

Protein 1	Description	Protein 2	Description	EOCT	EOCM
PEX12	PEX12 Peroxisome assembly protein 12	PEX11B	PEX11B Peroxisomal membrane protein 11B	1896.6	3534.5
PEX11B	PEX11B Peroxisomal membrane protein 11B	PEX13	PEX13 Peroxisomal membrane protein PEX13	1433.5	2671.4
PEX11B	PEX11B Peroxisomal membrane protein 11B	PEX11A	PEX11A Peroxisomal membrane protein 11A	1652.5	1564.5
GDI1	GDI1 Rab GDP dissociation inhibitor alpha	RND1	RND1 Rho-related GTP-binding protein Rho6	39.6	1507.3
FHL1	FHL1 Isoform 2 of Four and a half LIM domains protein 1	FHL5	FHL5 Four and a half LIM domains protein 5	3307.4	1408.8
F2	F2 Prothrombin (Fragment)	FGB	FGB Fibrinogen beta chain	93341.1	1104.8
ATF3	ATF3 Isoform 1 of Cyclic AMP-dependent transcription factor ATF-3	ATF4	ATF4 Cyclic AMP-dependent transcription factor ATF-4	4373.6	946.4
ZNF331	ZNF331 Zinc finger protein 331	ZNF3	ZNF3 Zinc finger protein 3	3492.4	755.7
B2M	B2M Beta-2-microglobulin	HLA-DPA1	HLA-DPA1 Major histocompatibility complex, class II, DP alpha 1	1351.0	575.4
B2M	B2M Beta-2-microglobulin	TRBC1	TRBC1 T-cell receptor beta chain C region	1164.4	496.0
DPF2	DPF2 Zinc finger protein ubi-d4	ZNF3	ZNF3 Zinc finger protein 3	5236.2	31.5
E2F1	E2F1 Transcription factor E2F1	SMAD3	SMAD3 Mothers against decapentaplegic homolog 3	2032.7	24.1

Continued on Next Page...

Table 4.1 – Continued

Protein 1	Description	Protein 2	Description	EOCT	EOCM
CD4	CD4 T-cell surface glycoprotein CD4	IGHG1	IGHG1 IGHG1 protein	1696.6	20.1
CSK	CSK Tyrosine-protein kinase CSK	CALM1	CALM3;CALM1;CALM2 Calmodulin	1430.7	16.94
HDAC1	HDAC1 Histone deacetylase 1	ZEB1	ZEB1 Zinc finger E-box-binding homeobox 1	1141.9	13.5
CDK2	CDK2 Cell division protein kinase 2	MCM2	MCM2 DNA replication licensing factor MCM2	2623.9	5.0

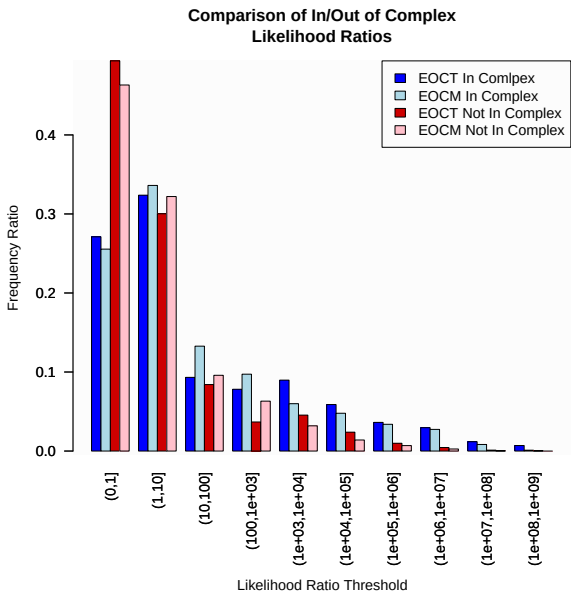
4.3 Cluster Analysis

The PIPs predictor identifies potential interactions based on the similarities between different types of evidence, whether they are co-expressed, have known orthologs or have similar annotations. As a result certain protein interactions, such as those that are in the same complex, are going to be easier to predict than others.

The HPRD contains annotations describing proteins that are co-complexed as well as which of the co-complexed proteins are known to interact. Figure 4.2 is a plot of the proportion of protein pairs that are co-complexed with likelihood ratios greater than or equal to defined thresholds. This shows that protein pairs that are part of the same complex tend to have a higher likelihood ratio than protein pairs that are not annotated as being part of the same complex.

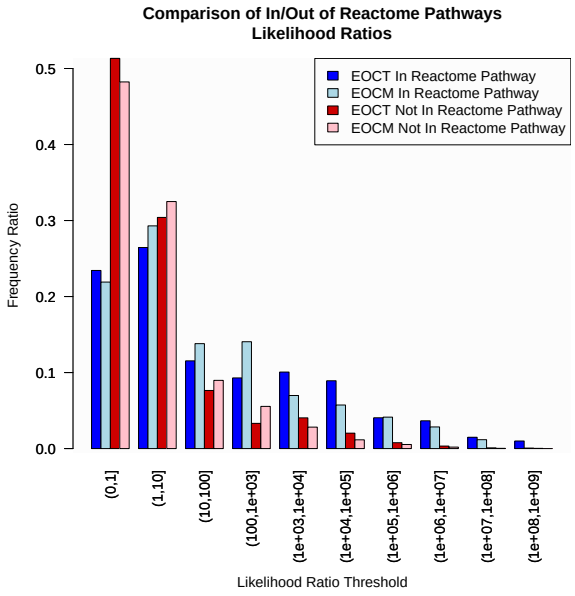
Figure 4.2 also shows the proportion of protein pairs that are not co-complexed

Figure 4.2: Plot of the proportion of interactions that are co-complexed/non co-complexed that have a likelihood ratio greater than of equal to a set threshold. Complex data is provided by the HPRD (Keshava Prasad et al., 2009).



and discretised into the same bins. There is a significant difference in the distribution of the likelihood ratios assigned to co-complexed versus non co-complexed protein pairs with a p-value estimated using the KS-test for scores assigned by either EOCT or EOCM of $< 2 \times 10^{-16}$. This is expected from the data as protein pairs that are part of the same complex are most likely to be co-expressed, have orthologous interactions, similar Biological Process annotations and significant co-occurrence of domains. The Transitive and Clustering modules would also act to enhance such interactions due to the highly similar evidence for a set of co-complexed proteins.

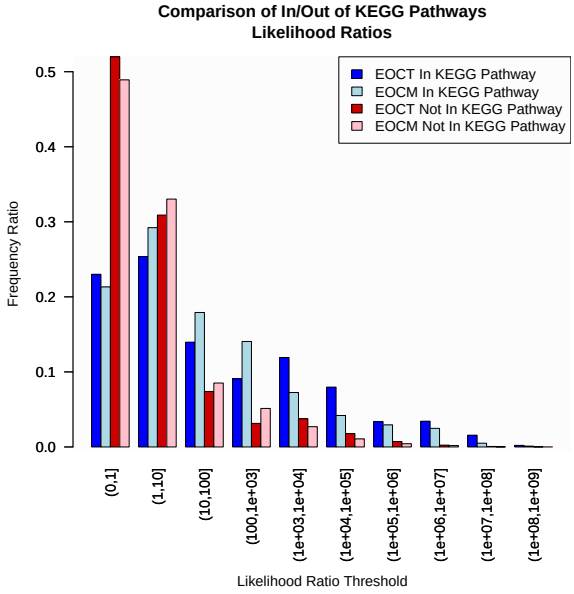
Figure 4.3: Ratio of protein pairs that are part of the same Reactome Pathway and between set likelihood ratio thresholds (blue), plotted along with protein pairs discritised by their likelihood ratio.



These results are not unique to proteins pairs that are co-complexed. Similar plots (Figure 4.3, Figure 4.4) and p-values (both $< 2 \times 10^{-16}$) are obtained for proteins that are part of the same Reactome (Matthews et al., 2009) or KEGG (Kanehisa, 2002; Kanehisa et al., 2006) pathways. It is therefore possible to identify clusters of proteins that have a significantly enriched set of interactions and are

therefore more likely to co-complex or be part of similar pathways based on the calculated likelihood ratios.

Figure 4.4: Ratio of protein pairs that are part of the same KEGG Pathway and between set likelihood ratio thresholds (blue), plotted along with protein pairs discretised by their likelihood ratio.



4.3.1 Identification of Significant Sets of Proteins

The previous section identifies that it is possible, given a cluster of proteins to determine if there is a significant enrichment of protein-protein interactions. Therefore, if given a cluster of genes or proteins are identified from a microarray next generation sequencing experiment or a proteomic study, it is possible to identify whether that set is significantly enriched for protein-protein interactions.

To analyse whether scores assigned to clusters of proteins were significant, such as co-complexed or present in the same biological pathway, randomly selected clusters of proteins over different cluster sizes were generated and scored based on the sum of the square of the likelihood ratios between proteins in the set (Equation 4.3.1).

Sum of Squares was chosen as it would easily highlight clusters of proteins that had high likelihood ratios.

$$S_s = \sum_{i \in I} LR_i^2$$

Equation 4.3.1: Sum of Squares

Where I is the set of all possible interactions within a set of proteins.

The cluster sizes ranged from 3 through to 1000, with randomly selected proteins and each selection was repeated 10,000 times for each cluster size. Figure 4.5 shows the cumulative plot of the frequency of clusters of varying sizes that had scores greater than or equal to increasing threshold scores. From this it is possible to estimate the 95% confidence interval and therefore apply that to known clusters of proteins derived from biological complexes, Reactome or KEGG annotations. Figure 4.6 shows a histogram of the scores for clusters of proteins based on Reactome

Figure 4.5: Sum of squares scores (EOCM and EOCT respectively) for randomly generated clusters of a predefined number of proteins. Marked in grey is the 95% confidence interval.

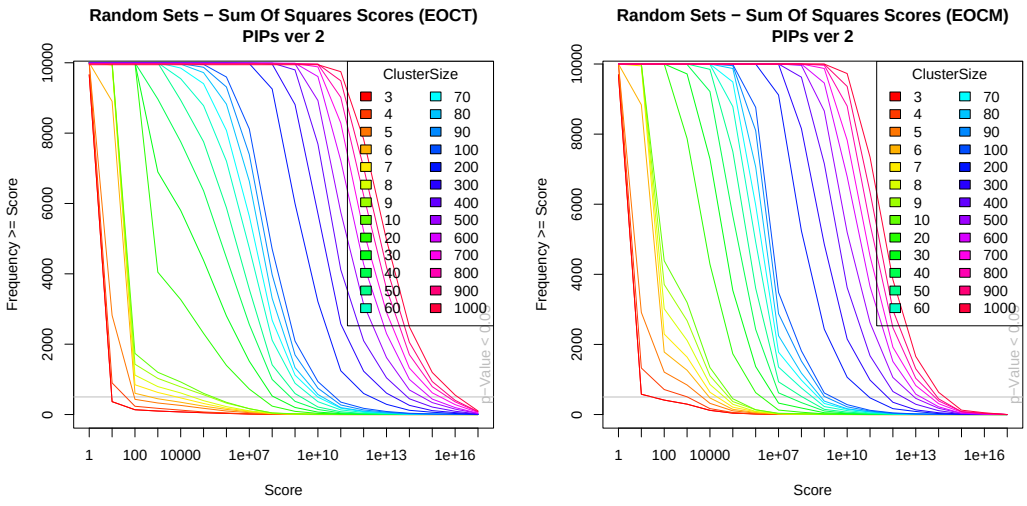


Figure 4.6: Histogram of clusters of proteins based on their membership to Reactome pathways. The clusters of proteins are scored based on the sum of squares scores for the likelihood ratios between all of the proteins within the clusters. In red are clusters of proteins where the p-value is less than or equal to 0.05.

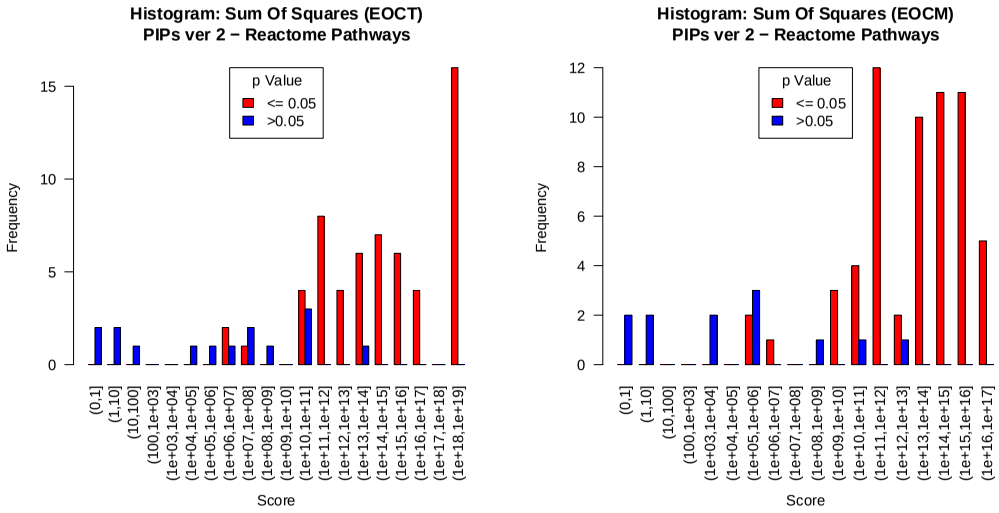
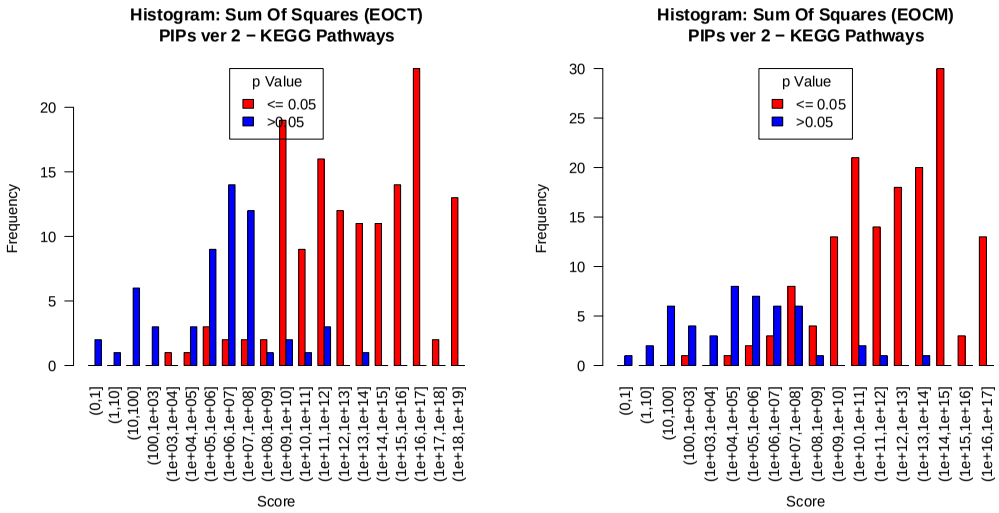
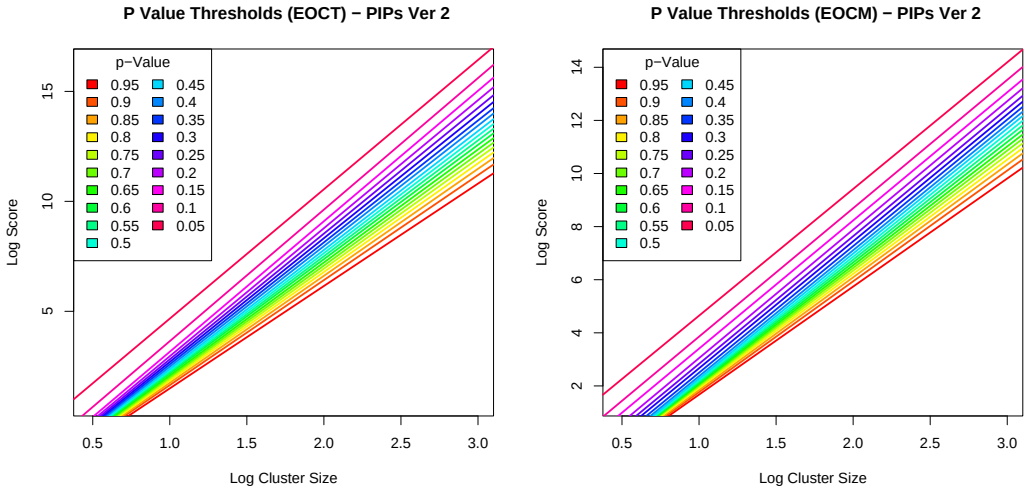


Figure 4.7: Histogram of clusters of proteins based on their membership to KEGG pathways. The clusters of proteins are scored based on the sum of squares scores for the likelihood ratios between all of the proteins within the clusters. In red are clusters of proteins where the p-value is less than or equal to 0.05.



annotations and Figure 4.7 is a histogram of complexes based on KEGG annotations. Both Figure 4.6 and Figure 4.7 show that it is possible to confidently identify significant clusters of proteins based on likelihood ratios calculated using PIPs.

Figure 4.8: The lines represent the p-value for given cluster sizes and scores for EOCT (left) and EOCM (right). The x axis is the log of the number of proteins in the cluster and the y axis is $log(S_s)$ for a given cluster.



For the random cluster scores, as shown in Figure 4.8 there is a linear log-log relationship between the cluster size and the score assigned to the cluster. It is then possible to estimate a p-value for any given cluster of proteins based on the cluster size and the cluster score. Table 4.2 shows the gradients (m) and y axis intersect points (c), from these it is possible to estimate the p-value for a cluster for any given size or score. Figure 4.9 shows linear regression lines for fitting the m and c values to given p-values. Linear regressions have been used to simplify the problem rather than using a single complex polynomials. Equation 4.3.2 can be derived from the two straight line equations to calculate the p from m and p from c.

$$p = \frac{fx + h - y}{dx + g}$$

Equation 4.3.2:

where p is the p-value that is to be calculated, x is the log of the cluster size and y is the log of the cluster score. The values d, f, g and h are a constant dependent

Table 4.2: Points for determining the p-value for a cluster of proteins. Where m is the gradient of the line and c is the y axis intersect.

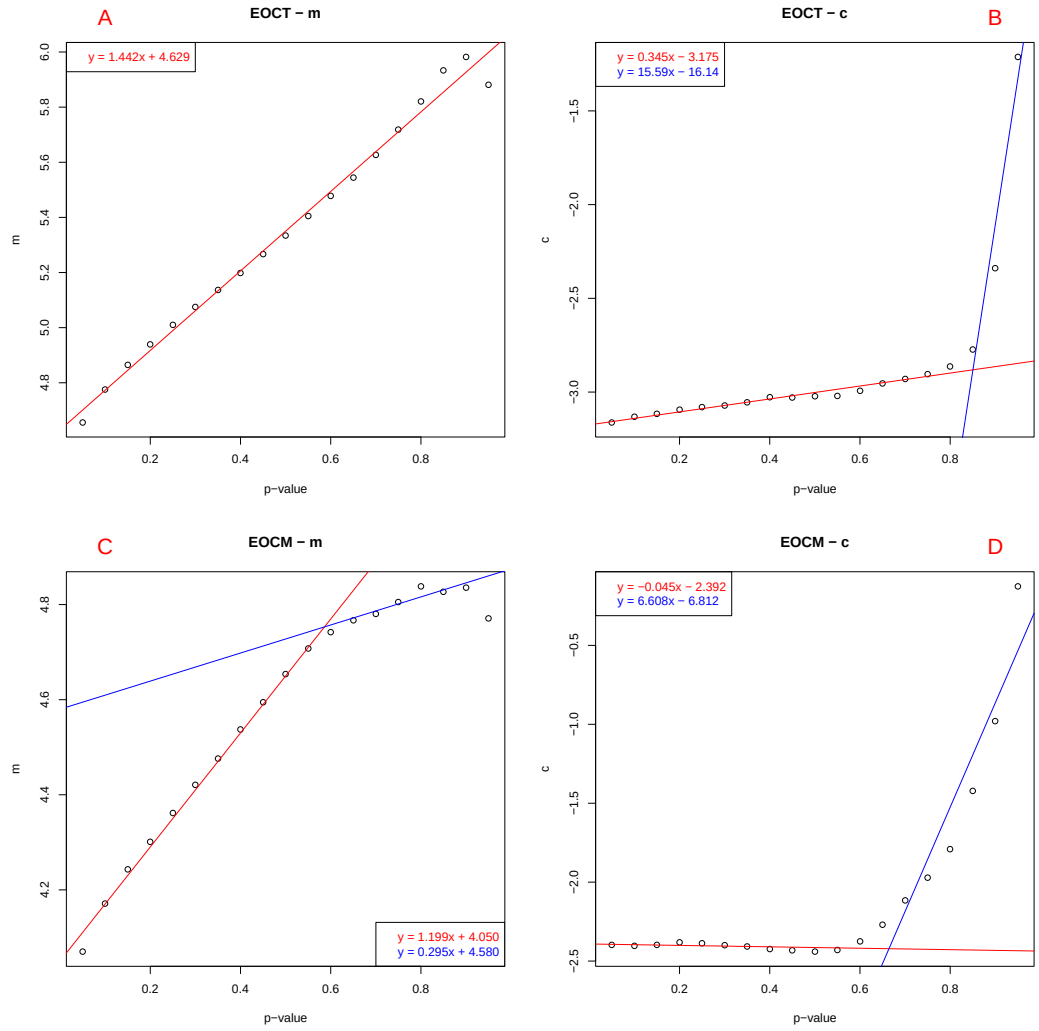
p Value	EOCT		EOCT	
	m	c	m	c
0.05	4.66	-3.16	4.07	-2.40
0.1	4.78	-3.13	4.17	-2.40
0.15	4.87	-3.12	4.24	-2.40
0.2	4.94	-3.09	4.30	-2.38
0.25	5.01	-3.08	4.36	-2.39
0.3	5.08	-3.07	4.42	-2.40
0.35	5.14	-3.05	4.48	-2.41
0.4	5.20	-3.03	4.54	-2.42
0.45	5.27	-3.03	4.60	-2.43
0.5	5.33	-3.02	4.65	-2.44
0.55	5.41	-3.02	4.71	-2.43
0.6	5.48	-2.99	4.74	-2.38
0.65	5.54	-2.95	4.77	-2.27
0.7	5.63	-2.93	4.78	-2.12
0.75	5.72	-2.9	4.81	-1.97
0.8	5.82	-2.86	4.84	-1.79
0.85	5.93	-2.77	4.83	-1.42
0.9	5.98	-2.34	4.84	-0.98
0.95	5.88	-1.21	4.77	-0.13

Table 4.3: Fixed values for calculating the p-value

	EOCT		EOCM	
	1	2	1	2
d	1.442	1.442	1.199	0.295
f	4.629	4.629	4.050	4.580
g	0.345	15.59	-0.045	6.608
h	-3.175	-16.14	-2.392	-6.812
p-value range	0 to 0.8	0.8 to 1.0	0 to 0.65	0.65 to 1.0

on which p-value is being calculated; see Table 4.3 for actual values and over which p-value ranges they apply.

Figure 4.9: Mapping p-values to gradient and y-axis intersect points using linear regression of EOCT and EOCM scoring methods. Where y is c or m and x is the given p-value.



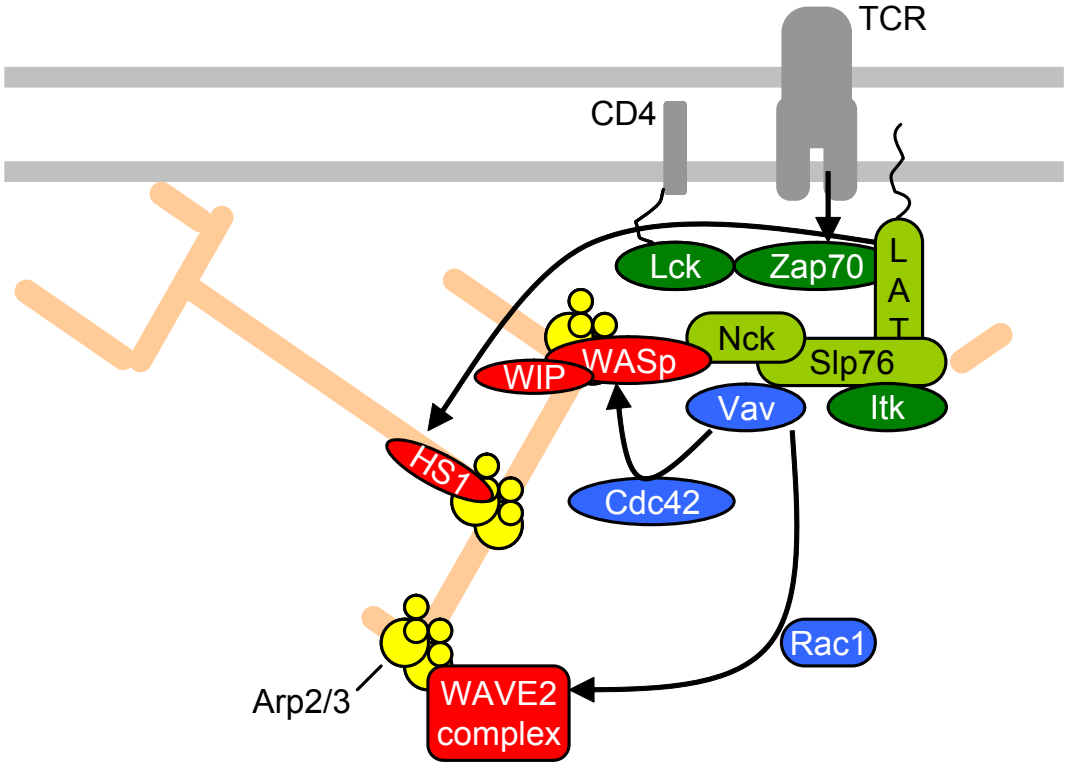
4.4 Biologically Significant Predictions

This section describes several biologically interesting groups of proteins that have been identified by the PIPs 2 predictor. The groups analysed include the T-Cell signalling pathway, the proteasome and the nuclear import and export complex. These proteins were selected by manually reading through the predictions for novel interactions and then identifying whether then surrounding interactions would indicate whether this is a viable interaction.

4.4.1 T-Cell Signalling Pathway

The T-cell receptor is present on the surface of T cells that recognise antigens presented by other cells (Burkhardt et al., 2008). Upon receiving a signal from the

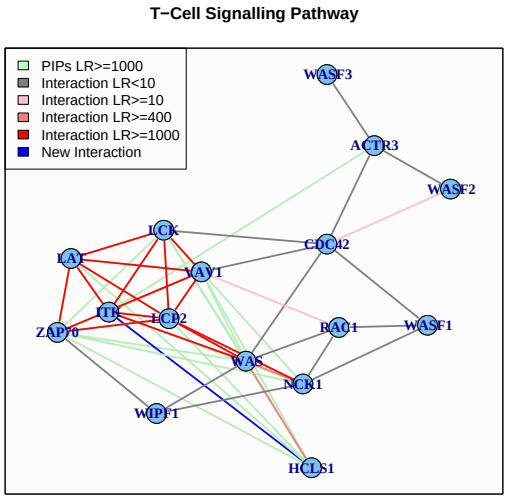
Figure 4.10: T-Cell Receptor (TCR) signalling pathway involved in the remodelling of the actin cytoskeleton (orange). Adapted from Burkhardt et al. (2008). The two horizontal grey lines represent the plasma membrane.



T-Cell Receptor (TCR), the T-Cell remodels its actin, thus allowing the T-cell to migrate from the blood into the tissue. Figure 4.10 (Burkhardt et al., 2008) shows the signalling cascade from the TCR to initiate actin remodelling and the protein-protein interactions that have to occur to allow for the restructuring of the T-Cell to penetrate into the tissue.

Figure 4.11 shows the predicted interaction network, as calculated by PIPs, between proteins involved in the TCR signalling cascade. The interaction that is highlighted in blue is between HCLS1 (also known as HS1 in Figure 4.10) and ITK. This has been validated as a true interaction (Carrizosa et al., 2009) as HSCL1 and ITK interact to recruit actin to the TCR and initiate the remodelling of the actin cytoskeleton allowing the migration of the T-cell into the tissue.

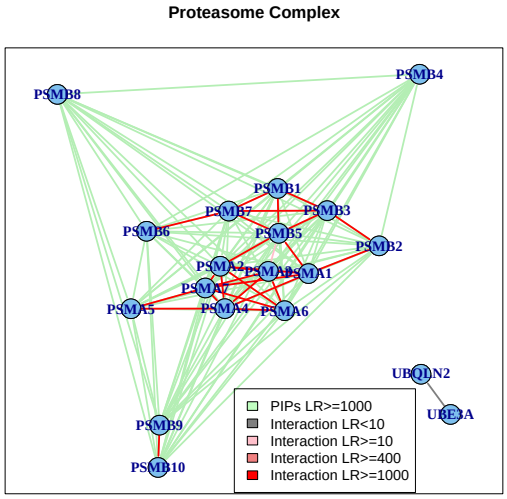
Figure 4.11: Predicted interactions between proteins involved in the T-Cell Signalling pathway as highlighted in Figure 4.10. Red lines are known interactions, where the gradient from grey through to red is dependent on the calculated likelihood ratio as determined by the PIPs predictor. Lines that are highlighted in green are predicted interactions that have a likelihood ratio (EOCT or EOCM) ≥ 1000 and the line highlighted in blue is between ITK and HCLS1 which has a likelihood ratio of 34202.5 and 6272.2 (EOCT, EOCM respectively) and has been validated as a true interaction (Carrizosa et al., 2009).



4.4.2 Proteasome Complex

Interactions within the proteasome complex were not included as part of the training and testing of the PIPs predictor as they had been assigned to dataset 6 (see Figure 3.2). However, as Figure 4.12 shows, the PIPs identifies all of the interactions in the HPRD (Keshava Prasad et al., 2009). This is not a challenging prediction as the proteasome is a large and well studied complex and is therefore well annotated, plus conserved across yeast, worm and fly. The genes encoding for the proteasome have similar gene expression patterns inferring strong support for the predicted interaction. This also explains the large connectivity between all of the proteasomal subunits.

Figure 4.12: The proteins known to make up the Proteasome complex and their predicted interactions. The interactions that are highlighted in grey through to red are known interactions where the colour indicates the calculated likelihood ratio (EOCT and EOCM). Edges that are highlighted in green indicate predicted interactions with likelihood ratios ≥ 1000 (EOCT and EOCM).



4.4.3 Nuclear Import and Export

Access into the nucleus is regulated via the nuclear pore complex through the nuclear membrane. Soluble small molecules are able to freely flow in and out, but larger molecules and proteins require import/export with a licensing factor. The network of interactions that occur between the proteins of the import and export pores are shown in Figure 4.13 with interactions of interest are highlighted with purple edges. The predicted interaction between Exportin (XPO1) and RanBP3 is known to occur for the shuttling of cargo from within the nucleus to the cytoplasm (Lindsay et al., 2001); however it was not included in the training and test set, but was identified by the predictor. A similar interaction was also predicted to occur between the proteins Importin β1 (KPNB1) and RanBP1, although with a likelihood ratio of 901.85 and

Figure 4.13: PIPs predicted nuclear import/export pore related proteins. Grey through to red edges indicate known interactions (Keshava Prasad et al., 2009) with the gradient depending on the calculated likelihood ratio. Green indicates interactions predicted by PIPs, but not present within the database, the thin lines indicate predicted interactions with LR $\geq$ 1000 (EOCT or EOCM). The purple edges are interactions of special interest.

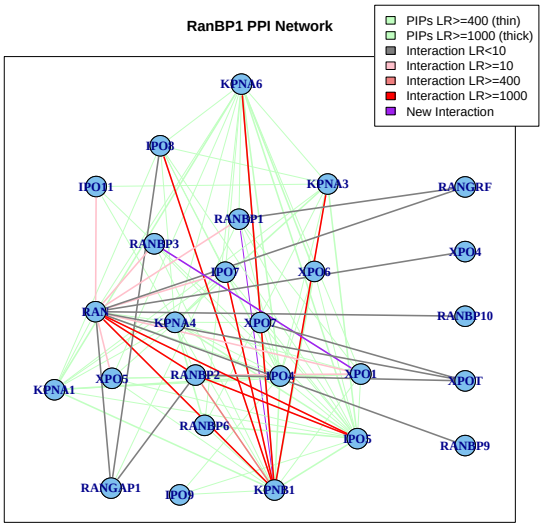


Table 4.4: Network Sizes

	Intersect	Union
Number of Proteins	5553	12633
Number of Interactions	94507	310894

954.35 (EOCT and EOCM respectively), therefore the posterior odds ratio is less than 1. The interaction between KPNB1 and RanBP1, which allows for the import of molecules from the cytoplasm into the nucleus (Lonhienne et al., 2009), has not been annotated within the HPRD database (Keshava Prasad et al., 2009).

With the predicted interactions between XPO1 interacting with RanBP3 and KPNB1 interacting with RanBP1 already published in the literature this highlights one of the problems of curated databases. Within the literature there are many interactions that are known to occur, but they are not annotated within managed databases, making them inaccessible to computational analysis and the analysis of the completion of the interactome.

4.5 Network Analysis and Co-Localisation

Two different predicted interaction networks are considered within this section. The first set is the intersect (LR1000i), this is derived from interaction predictions where both EOCT and EOCM give a likelihood ratio ≥ 1000. The second network is the Union set (LR1000u), this is where the interaction predictions have a likelihood ratio ≥ 1000 from either EOCT or EOCM. Table 4.4 shows the sizes of the two networks.

Figure 4.14 and Figure 4.15 show the intersect and union networks (LR1000i and LR1000u respectively). Although on their own the figures of the two networks are not hugely informative, calculated properties of the networks can be used to

characterise their structure, such as the clustering coefficient and degree. The degree of a node in a network is the number of edges that it has to other nodes in the network, in this case each node represents a protein and the edges are interactions with other proteins. The clustering coefficient is a measure of how interconnected a network is.

The Figure 4.16A is a log-log plot of the degree distribution versus the degree of the nodes, suggesting that both the intersect and union networks are either scale

Figure 4.14: The Intersect of interactions predicted by PIPs where both EOCT and EOCM calculated a likelihood ratio of interaction ≥ 1000.

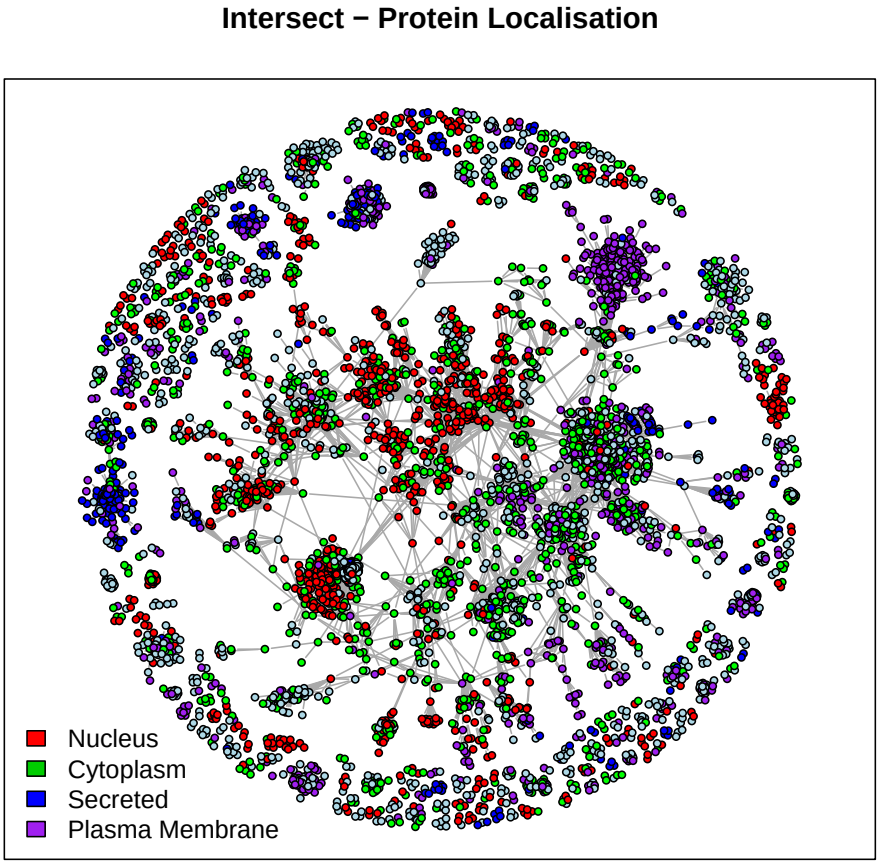


Figure 4.15: The Union of interactions predicted by PIPs where both EOCT or EOCM calculate a likelihood ratio of interaction ≥ 1000.

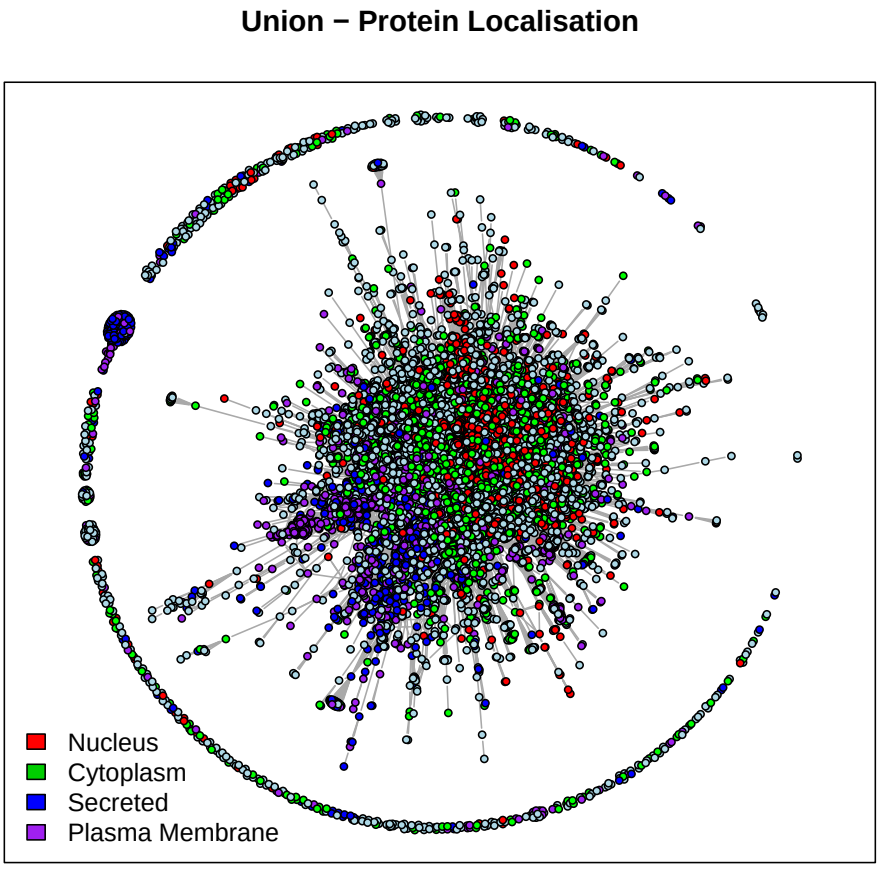
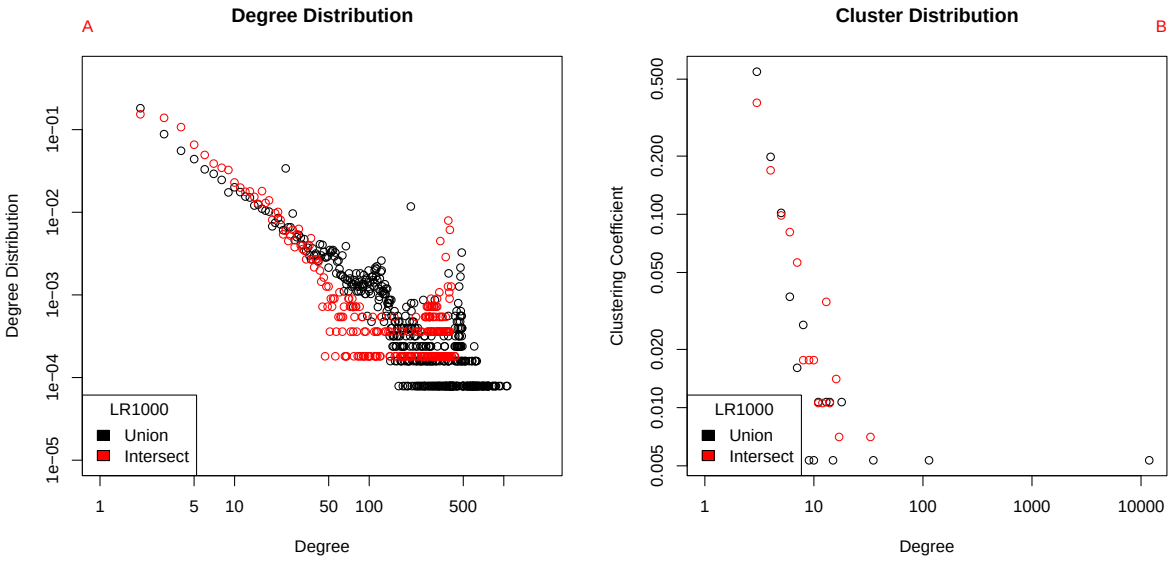


Figure 4.16: (A) Degree Distribution plot of the Union and Intersect EOCT and EOCM LR1000 sets. (B) Cluster coefficient plot versus the degree. The plots indicate that there is a hierarchical structural within the LR1000i and LR1000u predicted interaction networks.



free, or hierarchical, networks, as found previously numerous times in human and other organisms (Barabasi and Oltvai, 2004; Gandhi et al., 2006; Jeong et al., 2000; Wagner and Fell, 2001). However, as shown in Figure 4.16B which is a log-log plot of the cluster coefficient versus the degree which suggests that they are hierarchical and therefore identifying the presence of embedded modules, or sets of proteins, within the predicted networks that are more highly connected (Barabasi and Oltvai, 2004). Hierarchical networks have previously been found in protein interaction networks, such as those of metabolism (Ravasz et al., 2002). The largest connected component within the LR1000i set contains 3466 proteins and 85,423 edges and therefore represents the majority of the network.

The set of network diagrams (Figure 4.14 and Figure 4.15) identify proteins that are colocalised within the cell. The identifications are made based on The Gene

Ontology (GO) annotations (Ashburner et al., 2000) associated to the proteins. Due to the hierarchical nature of the GO, terms assigned to proteins were used as the starting points to trace back to the relative localisation term. Figure 4.17 shows the relationship between interacting proteins between two localisations within the cell based on the LR1000i set. The coefficient calculated for each square uses Equation 4.5.1 (Yook et al., 2004).

$$l(\lambda,\theta) = \frac{L^{\lambda,\theta} + L^{\theta,\lambda}}{L^{\lambda} + L^{\theta}} \equiv \frac{2L^{\lambda,\theta}}{L^{\lambda} + L^{\theta}} \equiv \frac{2L^{\lambda,\theta}}{L^{\lambda} + L^{\theta}}$$

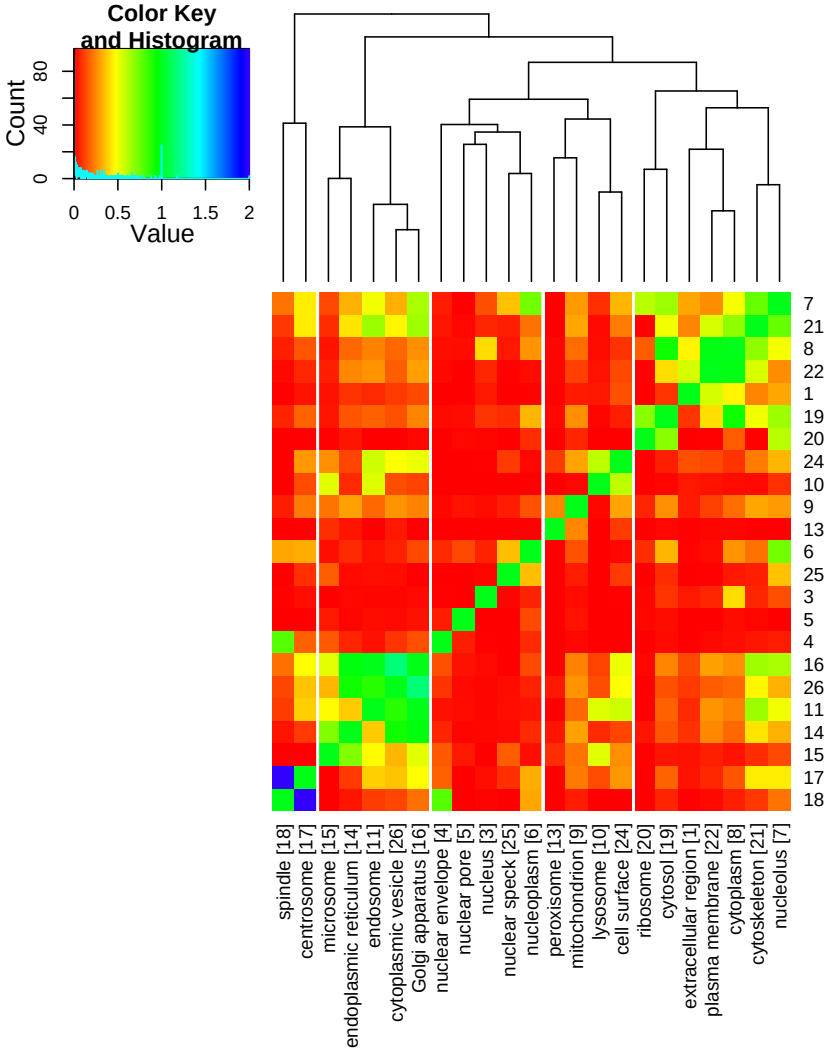
Equation 4.5.1:

where $l(\lambda,\theta)$ is the coefficient of proteins within one cellular compartment (λ) interacting with proteins of a second compartment (θ). $L^{\theta,\lambda}$, is the number of interactions between proteins of compartment λ that interact with proteins in compartment θ. L^{λ} is the total number of interactions that occur between proteins in compartment λ. Therefore, if λ and θ are the same compartment, then the output value is 1. If the two sets of proteins do not interact with each other then the output score is 0 and if there are more interactions between the compartments than within the compartments then the score is greater than 1.

The number of interactions between compartments is low with the majority < 0.5. There are also some compartments where there are more interactions between the two compartments than there are within the two compartments individually (Cytoplasmic vesicle and Golgi apparatus and the centrosome and spindle where the coefficient of interaction is > 1). Compartments that have a raised number of interactions between two components include those that are part of the secretory pathway, such as the endoplasmic reticulum, golgi apparatus, microsome, endosomes and cytoplasmic vesicles. There are also increased numbers of links between the nucleoplasm, nucleolus, cytoplasm, cytosol, ribosomes, cytoskeleton and plasma membrane.

The prediction of more proteins interacting between the centrosome and the spindle rather than between proteins of each of the two compartments is not surprising

Figure 4.17: Measure of the coefficient of interactions (Yook et al., 2004) between proteins annotated as present within separate compartments of the cell as defined by the assigned GO terms (Ashburner et al., 2000) to the proteins. The numbers on the y-axis correspond the numbers in square brackets along the x-axis indicating the compartment of comparison. The colours of red through to blue indicate an increasing number of interactions between the compartments in ratio to the number of interactions within each of the compartments. Green indicates that there are the same number of interactions between compartments as they are within the compartment. The compartments have been clustered based on a hierarchical clustering (top of the diagram) of the coefficients of interactions between all compartments. The histogram is the coefficients of interactions represented in the matrix.



due to interactions during cell division. There are also an increased number of interactions occurring between the spindle and the nuclear envelope. The interaction between the spindle and the nuclear envelope could be inferred due to the use of the orthology module because in yeast the nuclear envelope does not break down during mitosis and the spindle body is embedded within the envelope itself (Castillo et al., 2002).

On clustering the calculated coefficients using a hierarchical clustering method (R, hclust) it is possible to identify common compartments that are highly interconnected (Figure 4.17). The groups of compartments that are identified are:

1. Centrosome and Spindle,
2. Endosome, Golgi Apparatus, Cytoplasmic Vesicle, Endoplasmic Reticulum and Microsome,
3. Nucleus and Nuclear Pore, Nuclear Envelope, Nucleoplasm and Nuclear Speck,
4. Lysosome and Cell Surface, Mitochondrion, Peroxisome,
5. Cytosol and Ribosome, Nucleolus and Cytoskeleton, Extracellular Region, Cytoplasm and Plasma Membrane.

In group 4 the predictions suggest that there are few interactions that occur between proteins that are located within these compartments and other compartments within the cell. The nucleus is the one exception to that where some of its proteins also interact with proteins in the cytoplasm, although these proteins are more likely to exist in both compartments rather than actual interactions between the compartments.

In group 2, all of these compartments are part of the secretory pathway. Therefore, these proteins are likely to be trafficked through each of these compartments. Proteins that are located within these compartments are therefore more likely to come in contact and interact, especially if they are signalled for trafficking. The interactions can occur between the proteins that are being exported, or transported to the cell surface along with other proteins that are involved in the trafficking, but are not destined for export or localisation to the cell surface.

Similar to group 2, the proteins in group 5 are also likely to share common compartment interactions. Proteins that are present within the cytoplasm are likely to interact with proteins that are in the plasma membrane as are proteins within the plasma membrane to interact with proteins that are secreted from the cell (Extracellular Region).

4.6 Validation of Predicted Protein-Protein Interaction

The challenge with computational predictors is to obtain independent validation of the predictions that have been made. In the case of PIPs it is possible to use blind test sets to assess the quality of the predictions that have been made. However, being able to test the predictions provides further proof that the predictor is performing correctly and generating predictions that are correct. Ideally all predicted protein-protein interactions should be experimentally tested, but this is impractical. A more pragmatic approach is to select number of highly probable predicted interactions.

To test whether the interaction predictions can occur it is possible to use co-immunoprecipitation. Co-immunoprecitpitation requires that both proteins have a corresponding antibody that can bind specifically to the protein. The HPR (Human Protein Atlas Antibodies, release 6.0) has tested 11,132 antibodies that match to at least one protein within the PIPs database. In total the 11,132 antibodies map to 15,274 distinct proteins within the PIPs database, of which 10,419 are part of the informative protein set. There are 34,409 protein pairs where both proteins have a matching antibody and have EOCT and EOCM likelihood ratios ≥ 1000. If these 34,409 protein pair are filtered by known interactions, predicted interactions from OPHID (now i2d) (Brown and Jurisica, 2005) and genetic interactions from BioGRID (Stark et al. 2006), this leaves 17,063 protein pairs that are predicted to be more likely to interact than to not interact that could be tested. The prior odds were set at $\frac{1}{1000}$ due to the calculations being based on an incomplete network of interactions, if the prior odds is reduced to $\frac{1}{1200}$ (as calculated based on current known protein-protein interaction network sizes) this reduces the number of potential protein pairs to 12,331. From this it is possible to select a tractable number of interactions that can be experimentally validated.

4.7 Conclusion

This Chapter highlights that the protein-protein interaction predictions that are calculated by the PIPs predictor identify biologically significant protein pairs that are more likely to interact than to not interact. Interactions between protein pairs that are part of the same biological complex are more readily accessible via this

method.

With the use of Equation 4.3.2 and the values calculated in Table 4.3 it is possible to highlight sets of proteins that are more significantly linked. There are also compartments within the cell were proteins either interact between the compartments or are transported between compartments.

This Chapter also identifies 12,331 protein pairs (Section 4.6) that are likely to interact and have antibodies for both proteins therefore allowing the interaction to be experimentally verified.

Chapter 5

Web Services

Preface

This Chapter describes the PIPs webservice and the FuncPIPs webservice for making the predictions accessible to the FuncNet predictor.

5.1 Introduction

Predictions of protein-protein interactions can be made and new methods proposed, but the real benefit to biology is when the predicted interactions are made publicly accessible in a way that they can be searched, interrogated and used for further research. The development of interfaces to databases allowing users to access the wealth of information within these silos is vitally important so that the knowledge can be mined and used by others, but also to promote and ensure the further development of the resource.

Databases for protein interactions come in many different styles. Some, such as DIP (Salwinski et al., 2004), HPRD (Peri et al., 2004; Mishra et al., 2006), MPPI (Pagel et al., 2005) and IntAct (Kerrien et al., 2007a) provide simple tables of the

interactions contained within the databases. Other databases, such as STRING (Jensen et al., 2008) and MINT (Ceol et al., 2010) provide more interactive experiences with network representations of the interactions. For databases that provide predicted protein-protein interactions it is important to also provide access to the evidence used to make those predictions (Jensen et al., 2008; McDowall et al., 2009) (see Section 5.2) which is essential for users to be able to understand why a specific prediction was made. All databases provide the ability to download the interactions in a flat-file format allowing the interactions they contain to be compared and analysed by other users.

This chapter describes the web services that were created to allow access to the predictions made by the PIPs 1 predictor. The PIPs Web Service is a database front end that allows the users to query a protein to discover predicted protein-protein interactions and retrieve the evidence that the prediction was based on. FuncPIPs is a web service that allows programmatic access to the database for FuncNet, a predictor of functional interactions.

5.2 PIPs Webservice

http://www.compbio.dundee.ac.uk/pips

5.2.1 Database

The PIPs 1 database (Scott and Barton, 2007) contains details about 69,965 human proteins that have been imported from the IPI (Kersey et al., 2004) together with likelihood ratios for 17,643,506 protein pairs, of which 37,606 are more likely to interact than to not. For each protein pair a break down of the likelihood ratios is provided of the contribution made by each of the modules along with the supporting evidence for the calculated likelihood ratio. The evidence includes 5872 *S. cerevisiae*, 23,195 *C. elegans* and 27,629 *D. melanogaster* proteins that were analysed by InParanoid (Berglund et al., 2008) to identify orthologous protein pairs that were known to interact. Protein annotations, such as InterPro (Mulder et al., 2007) motifs and domains, post translational modifications and cellular localisation data are stored as well as Pearson's Correlation of coexpression between protein pairs. The database also provides links to external data sources such as RefSeq (Pruitt et al., 2007), UniProt (Consortium, 2008) and Entrez (Wheeler et al., 2008). Comparisons and links are also made to other publically available protein-protein interaction databases including HPRD (Peri et al., 2004; Mishra et al., 2006), DIP (Salwinski et al., 2004), BIND (Alfarano et al., 2005) and OPHID (Brown and Jurisica, 2005) for protein pairs represented in those databases.

The PIPs database is constructed on top of a Linux server running the MySQL database software and Apache/Tomcat for the web server. The front end utilises

Java Servlet Pages (JSP) to provide a dynamic and easy to navigate web interface.

5.2.2 PIPs Web Interface

The front page of the PIPs interface allows for simple searches of the database with IPI, UniProt or RefSeq identifiers for proteins or via a text search for keywords. There are options to alter the minimum threshold score. An Advanced Search option allows a query protein sequence to be searched against the PIPs database with MagicMatch (Smith et al., 2005), which returns an exact match. If there are no matching sequences, the user is given the option to perform a BLAST (Altschul et al., 1997) search. There is also the option for a Batch searching of the database if there are multiple protein sequence identifiers; this is limited to a maximum of 25 per request.

Figure 5.1 illustrates the results of a query for the protein IPI00016572 (SNRPG - Small nuclear ribonucleoprotein G). The Interaction Page lists the most likely proteins to interact with SNRPG which are ranked based on the calculated posterior odds ratio. For each interaction there is a breakdown of the contribution that is made by each of the modules, for example Figure 5.1 shows the predicted interaction of SNRPG with LSM8 and highlights that most of the contribution to the prediction was made by the expression and transitive modules, but there was a low contribution made by the orthology module. However, the interaction between SNRPG and SNRPD3 had strong contributions made by all modules. The "Evidence" column provides a link to the supporting evidence that was used by each module to calculate each likelihood ratio which contributes to the final posterior odds ratio. The "Database" column identifies if the protein pairs are present in

other databases; currently this includes BIND, DIP, HPRD and OPHID.

Figure 5.2 shows the Evidence for Interaction page for the predicted interaction between SNRPG and SNRPD3 shown in Figure 5.1. The page provides the evidence used by each module to calculate each likelihood ratio. It is broken down into 6 sections covering gene expression, orthology, domain annotations, post translational modifications, localisation and transitive interactions.

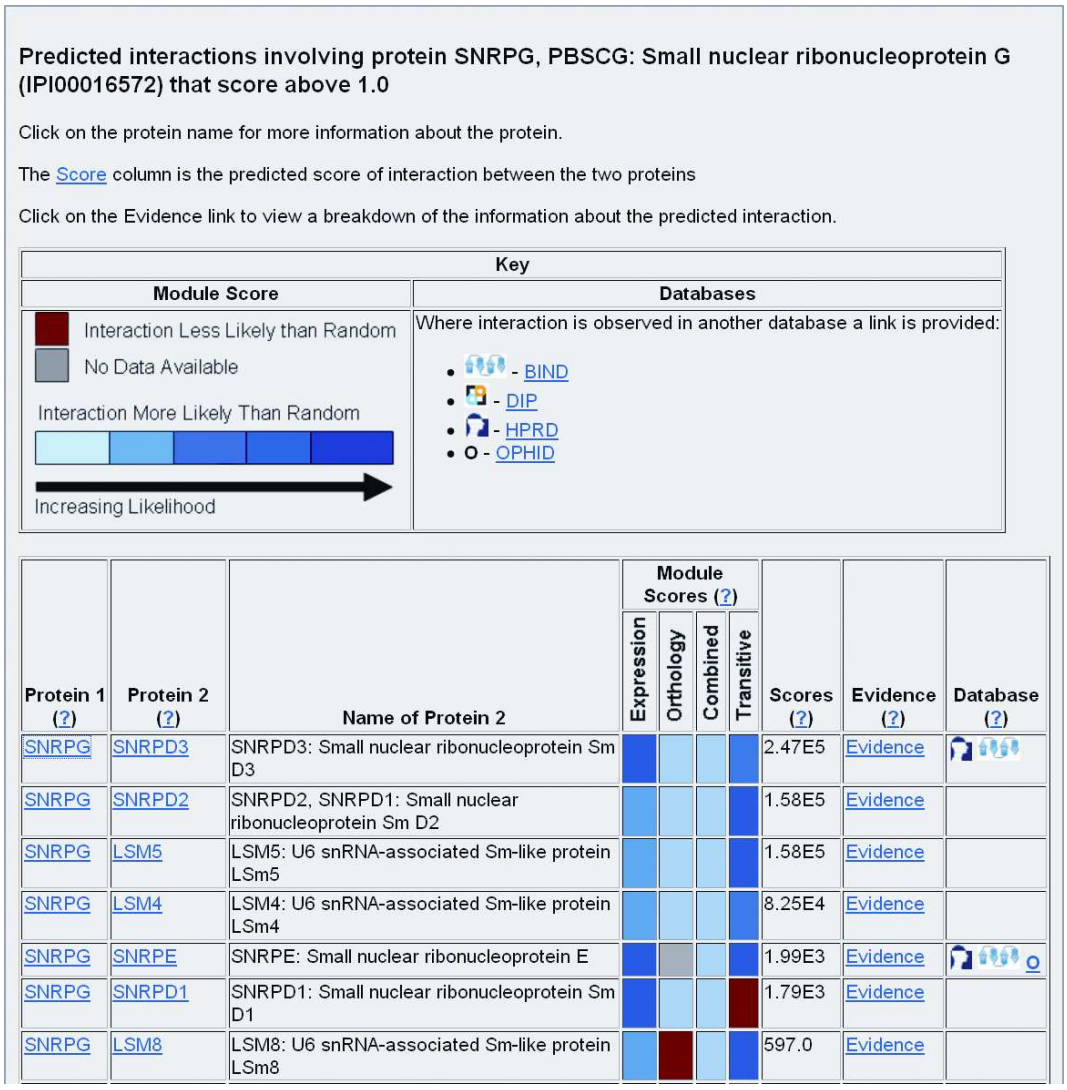
For each protein within the PIPs database there is a Protein Summary page. Figure 5.3 shows the summary page for the SNRPG protein. This page provides information about the number of interactions above set posterior odds values and the number of interactions known in other databases. There are also links to external databases for further information about the SNRPG.

Figure 5.4 illustrates the display of the predicted interactors of SNRPG via a Java applet that can be accessed from the Protein Summary page. Users are able to view the network of interactors that are predicted to interact with the query protein. The user is able to grow the network of interactions by selecting a protein of interest and clicking on the "Grow Network ..." option. Once the network has been generated the user is able to save the network as an image or download an adjacency list of the proteins so that they can be represented in an external application, such as Cytoscape (http://cytoscape.org) or Graphviz (http://www.graphviz.org).

5.2.3 Usage of the PIPs Webservice

Since the publication of the PIPs Webservice paper (McDowall et al., 2009), it has been cited 8 times. The number of unique page views of the PIPs Webservice in 2010 was $> 46,000$ with a total of $> 71,000$ page views. In total since June 2009,

Figure 5.1: Interaction Summary page for the protein IPI00016572 (SNRPG). The page shows the most probable protein interactions. There is a break down of the predictive features for each protein pair along with a link to further explore the evidence for the interaction.



Predicted interactions involving protein SNRPG, PBSCG: Small nuclear ribonucleoprotein G (IPI00016572) that score above 1.0

Click on the protein name for more information about the protein.

The Score column is the predicted score of interaction between the two proteins

Click on the Evidence link to view a breakdown of the information about the predicted interaction.

Key

Module Score

Interaction Less Likely than Random

No Data Available

Interaction More Likely Than Random

Increasing Likelihood

Databases

Where interaction is observed in another database a link is provided:

- BIND
- DIP
- HPRD
- O - OPHID

Protein 1 (?)	Protein 2 (?)	Name of Protein 2	Module Scores (?) Expression	Orthology	Combined	Transitive	Scores (?)	Evidence (?)	Database (?)
SNRPG	SNRPD3	SNRPD3: Small nuclear ribonucleoprotein Sm D3					2.47E5	Evidence	
SNRPG	SNRPD2	SNRPD2, SNRPD1: Small nuclear ribonucleoprotein Sm D2					1.58E5	Evidence	
SNRPG	LSM5	LSM5: U6 snRNA-associated Sm-like protein LSm5					1.58E5	Evidence	
SNRPG	LSM4	LSM4: U6 snRNA-associated Sm-like protein LSm4					8.25E4	Evidence	
SNRPG	SNRPE	SNRPE: Small nuclear ribonucleoprotein E					1.99E3	Evidence	O
SNRPG	SNRPD1	SNRPD1: Small nuclear ribonucleoprotein Sm D1					1.79E3	Evidence	
SNRPG	LSM8	LSM8: U6 snRNA-associated Sm-like protein LSm8					597.0	Evidence	

Figure 5.2: Evidence of Interaction Summary Page for the interaction between SNRPG and SNRPD3: (A) Sections Gene Expression and Orthology provide information about the correlation of coexpression between the two proteins and the orthology of the interacting pair. (B) Sections Domains, Post Translational Modifications and Localisations provide information about annotated protein domains present in both proteins, post translational modifications and the localisation of the proteins within the cell. (C) Section Transitive Score provides a list of transitive interactions between the two proteins with an integrated interaction score of > 0.025 for the Expression, Orthology and Combined modules. In total there are 236 predicted common interactors; the figure only shows the top six common interactors.

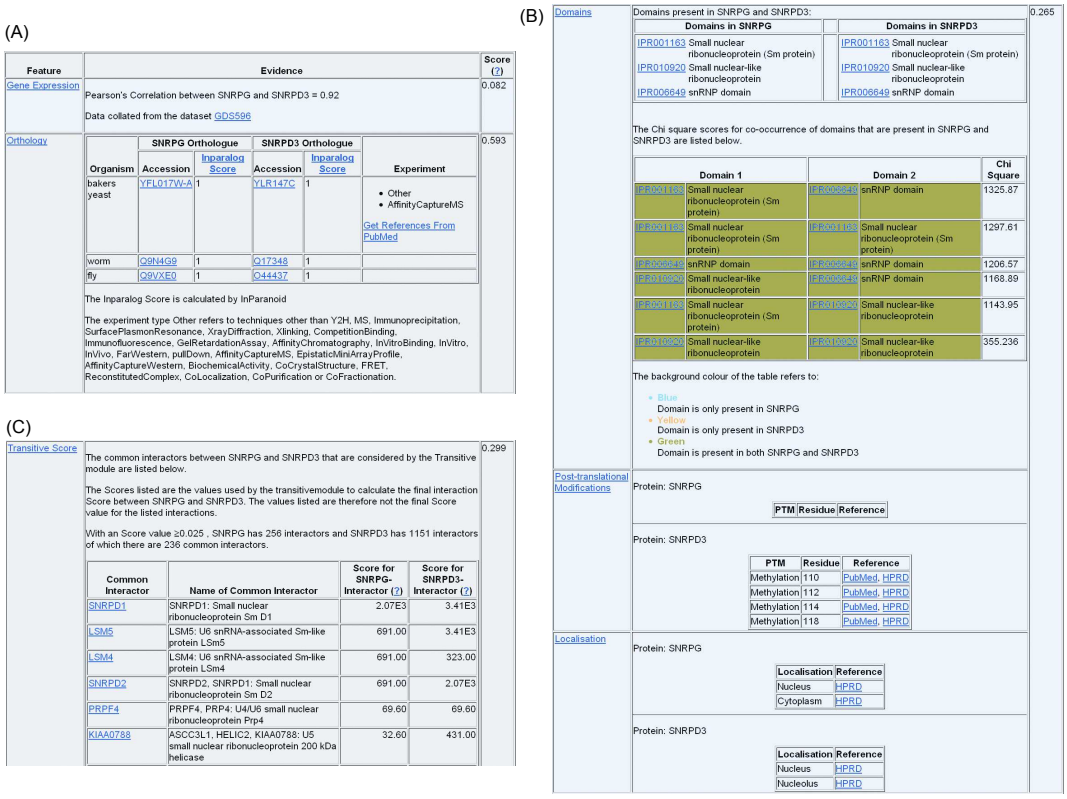
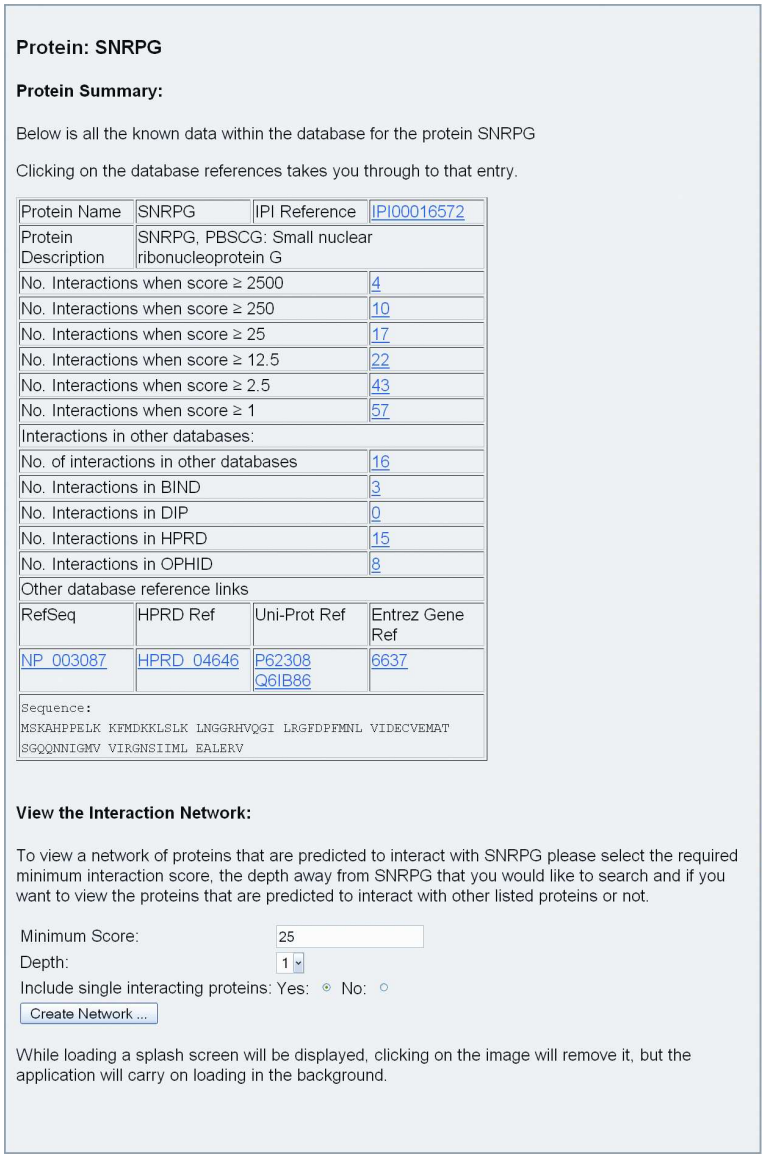


Figure 5.3: Protein Summary page for the protein SNRPG: Information about the selected protein including a breakdown of the number of predicted interactions at different threshold scores, the number of interactions in external databases. Links are provided for further information about the protein from RefSeq, HPRD, UniProt and Entrez.



Protein: SNRPG

Protein Summary:

Below is all the known data within the database for the protein SNRPG

Clicking on the database references takes you through to that entry.

Protein Name	SNRPG	IPI Reference	IPI00016572
Protein Description	SNRPG, PBSCG: Small nuclear ribonucleoprotein G		
No. Interactions when score ≥ 2500			4
No. Interactions when score ≥ 250			10
No. Interactions when score ≥ 25			17
No. Interactions when score ≥ 12.5			22
No. Interactions when score ≥ 2.5			43
No. Interactions when score ≥ 1			57
Interactions in other databases:			
No. of interactions in other databases			16
No. Interactions in BIND			3
No. Interactions in DIP			0
No. Interactions in HPRD			15
No. Interactions in OPHID			8
Other database reference links			
RefSeq	HPRD Ref	Uni-Prot Ref	Entrez Gene Ref
NP_003087	HPRD_04646	P62308 Q6IB86	6637

Sequence:
MSKAHPPELK KFMDKKLSLK LNGGRHVQGI LRGFDPFMNL VIDECVEMAT
SGQQNNIGMV VIRGNSIIML EALERV

View the Interaction Network:

To view a network of proteins that are predicted to interact with SNRPG please select the required minimum interaction score, the depth away from SNRPG that you would like to search and if you want to view the proteins that are predicted to interact with other listed proteins or not.

Minimum Score: 25
Depth: 1
Include single interacting proteins: Yes: No:
Create Network ...

While loading a splash screen will be displayed, clicking on the image will remove it, but the application will carry on loading in the background.

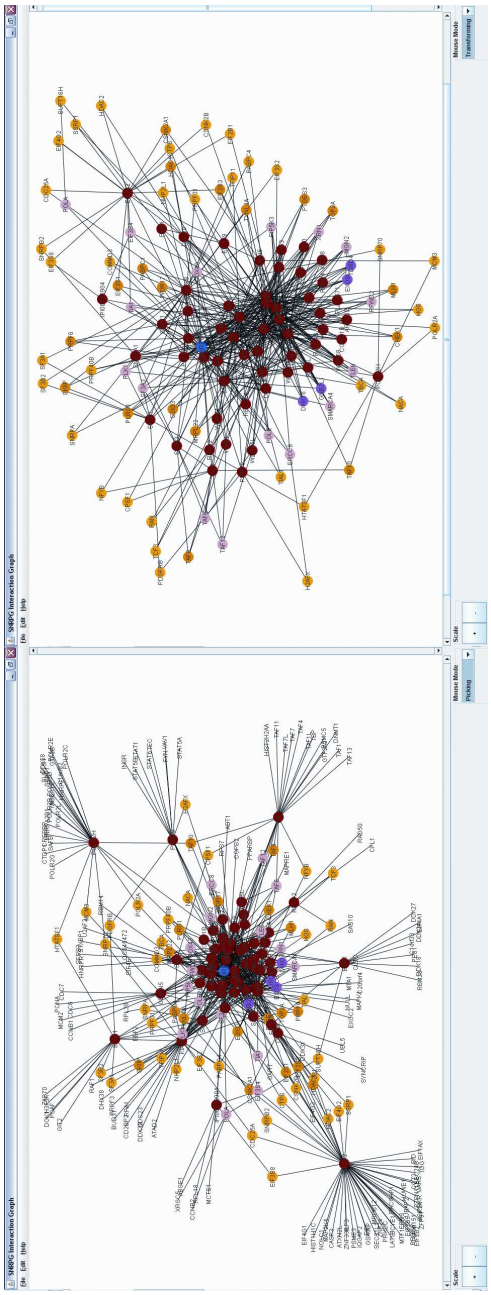


Figure 5.4: Network view of the predicted interactors of SNRPG: A Java application to view the local topology of the predicted protein-protein interaction network. Left: Highlighted in blue is the query protein (SNRPG) along with the predicted primary and secondary interactors. Proteins are highlighted dependent on the number of predicted interactors, yellow there are 2 interactors through to red with 5 or more interactors. Right: The network of predicted primary and secondary interactors of SNRPG (Blue), with all the interactors than have only a single predicted interaction removed.

when "Google Analytics" was used to monitor usage of the PIPs Webservice there has been a total of $> 98,803$ page views of which $35,401$ are unique. This indicates that there is a high level of interest in the webservice and the data that it provides.

5.2.4 Future Development

The new predictions that have been made (see Chapters 3 and 4) will be accompanied by updates to the PIPs web interface. The majority of the look and feel will remain the same to maintain a consistent experience for the user. The Evidence pages will change to a two tiered approach to reduce the time taken to populate a whole page. The first Evidence page will provide a summary of the evidence with the likelihood ratios calculated by each module, but this will link through to more detailed information about the calculation of the likelihood ratio. The Gene Expression will provide information about the source and the correlation along with dynamically generated graphs to illustrate the correlation of expression of the genes for the two proteins. The Orthology and Combined modules will remain the same, just separated into separate pages. The Transitive and Clustering modules will provide graphs of the predicted interaction of proteins that were considered in each module. The interaction viewer and the networks for the Transitive and Clustering evidence pages will be implemented as an applet that loads within the website rather than as a Web Start Java applet. Having the applet embedded within the web page as Javascript or Flash means that a program does not have to be downloaded by a user to view the network.

5.3 FuncPIPs

http://www.compbio.dundee.ac.uk/ws-pips/

FuncNet is a predictor of functional interaction of protein pairs developed as part of the ENFIN project. The FuncNet webserver aggregates numerous predictors of protein-protein interaction, both functional and physical, to assign the most likely functional interactions between two sets of proteins. Each server queried by FuncNet submits a request to the server to assign p-values to link proteins in a query set with those in a reference set.

The servers that are queried by FuncNet include:

- PIPs (McDowall et al., 2009)
- CODA (Reid et al., 2010)
- engineDB (Tulipano et al., 2007)
- GECO
- hiPPI
- iHOP (Fernandez et al., 2007)
- JACOP (Sperisen and Pagni, 2005)

For the protein-protein interaction predictions made by PIPs 1 to be included as part of FuncNet a p-value for the likelihood of an interaction between two proteins occurring by chance had to be calculated (Section 5.3.1). The web service had to be able to accept two sets (lists) of proteins and return the likely pairs of protein

interactions between the two sets along with the assigned p-value for the interaction (Section 5.3.2).

5.3.1 Calculating P-Values

The p-values were calculated by sampling 1,000,000 likelihood ratios out of the set of 155,000,000 predictions made by PIPs 1.

$$p = \frac{|\{x \in S : x \geq LR\}|}{|S|}$$

Equation 5.3.1:

where S is a set of 1,000,000 randomly sampled likelihood ratios ($|S| = 1,000,000$), x is an element of S and LR is the likelihood ratio of a preselected pair of proteins for which p is to be calculated. The 1,000,000 set of random likelihood ratios was sampled for each likelihood ratio where a p-value was to be calculated. To reduce the problem, p-values were only calculated for protein pairs that had a likelihood ratio greater than 1.0. It took just under 59 days to calculate p-values for the 17,643,506 protein pairs that have a likelihood ratio greater than 1. This was split into 505 jobs over the cluster of the College of Life Sciences, University of Dundee, so that it could be processed within 2 days.

5.3.2 Web Service

The database is hosted as a SOAP server that when queried returns the list of proteins between the two sets that have a precalculated p-value. The p-values are precalculated to minimise the response time. On test sets of 37 and 36 proteins it is possible to query all 666 potential pairs in under 1.7 seconds.

Chapter 6

Cross Organism Protein-Protein Interaction Prediction

Preface

This Chapter shows the application of the PIPs 2 predictor in other species and then analyses whether it is possible to train the predictor based on one organism and predict protein-protein interactions in a second organism.

6.1 Introduction

The PIPs predictor was built for the investigation of human protein-protein interactions, but not all research is done with human cells. To address the prediction of interactions in other organisms, the PIPs 2 framework and methodology can be applied in other organisms, provided relevant feature datasets and training examples are available. This allows for the PIPs predictor to be relatively species agnostic, requiring only minor tweaks to the thresholds or selection values dependent on the species used to train the predictor. As a proof of concept, there PIPs 2 predictor was built for three other organisms: *Saccharomyces cerevisae*, *Drosophila melanogaster*

and *Caenorhabditis elegans*, referred to as yeast, worm and fly, respectively, for the rest of the chapter.

6.1.1 Cross-Organism Model Prediction

Several model organisms and causative agents of disease are not included in many protein-protein interaction databases because too few interactions are known for these organisms. Such is the case for *Trypanosomas*, *Xenopus* and *Leishmania*.

There are several ways that predictions could be made for the likelihood of protein-protein interaction within a species that has no available database of protein-protein interaction:

Interologs Methods: Transfer known and predicted interactions from one organism to a second organism (Matthews et al., 2001).

Multi Species Prediction: Train a predictor on a multi-organism set of known protein interactions that interact based on a set of features across multiple species and use that model to predict new interactions within the same or other not too different organisms.

Cross Species Model Predictor: For an organism where there are few verified protein-protein interactions, but there is available feature datasets that can be analysed by a predictor, then a model could be trained using a second species that has a known set of protein-protein interactions and equivalent feature datasets. The trained model can then be applied to the feature datasets of the first organism to predict interacting pairs of proteins.

Transferring known protein-protein interactions from other organisms has been

used previously for making predictions of protein-protein interaction and functional inference of proteins in different organisms and has been shown to be successful (Matthews et al., 2001; Jensen et al., 2008; Wiles et al., 2010). However, making orthologous predictions limits the interactions that can be inferred. The limitation to such a method is firstly down to the interactomes being incomplete and secondly to the lack of a 1:1 mapping of proteins between all organisms. Therefore the more closely related two organisms are the larger the number of orthologues.

Martin et al. (2005) and Shaughnessy et al. (2008) have used sequence information where predictors have been trained and then applied in different organisms. In both papers Support Vector Machines were used to generate models based on features derived from the primary sequence of the protein pairs (interacting and non-interacting). Shaughnessy et al. (2008) found that they were able to build a predictor in one or multiple organisms and then accurately predict protein-protein interactions in those organisms; however when applying the predictor to the proteome of a new organism they were much less accurate. Martin et al. (2005) also noted that the accuracy that they achieved when training in human and testing in mouse is likely due to the close relationship in the proteomes, but when training in *H. pylori* and predicting in *E. coli*, their accuracy dropped.

The third method described above (Cross Species Model Predictor) would apply a model that has been trained in one organism and used to predict protein-protein interactions in a second organism based on equivalent evidence. This avoids the limitations of orthologous transfer of interactions as all proteins that have experimental evidence can be considered. In addition, because evidence from the second organism is considered, the overall accuracy of the final predictions might be greater

than the accuracy of a multi-species predictor.

The first half of the Chapter describes the development of the PIPs 2 predictor in other species. The second half of the Chapter describes the development of the PIPs 2 predictor for training based on the human training sets and then testing on a second species.

6.2 Development of PIPs 2 in Other Organisms

This section investigates how the PIPs framework can be applied to yeast, fly and worm as each have the required experimental data and a large number of known protein-protein interactions that have been experimentally determined and are present in repositories, such as IntAct, that can be used for training. The selection of organisms has been limited to those that are eukaryotic as PIPs was developed on human, which is eukaryotic, also there has been concerted efforts within the prokaryotes to develop protein-protein interaction predictors, for example STRING (Jensen et al., 2008). It should however, be theoretically possible to apply the PIPs 2 predictor to prokaryotic species as long as there is sufficient publicly available experimental data.

6.2.1 Methods and Data

Data and Module Construction

For each of the new species investigated, species-specific experimental data has been loaded into the database. The following datasets have been loaded for analysis by the PIPs 2 predictor:

Training/Test Sets The training and test sets were derived by the same method as that used for the Human PIPs predictor. The positive training sets for all species were derived from the IntAct database (Kerrien et al., 2007a). The negative protein pairs were selected at random and then known or predicted positives were filtered out. The selection of IntAct as the main source of protein-protein interaction annotations was instrumental to determine the feasibility of moving from one database to another and potentially of using IntAct, instead of the HPRD in later releases of the Human PIPs predictor. However, for this study, human protein-protein interaction annotations were derived from the HPRD (Peri et al., 2004; Mishra et al., 2006; Keshava Prasad et al., 2009) to match in with the current state of the PIPs 2 framework.

The protein pairs used for training were randomly split into 5 sets to allow for 5 fold cross validation by each of the module to access the accuracy of the module based on the training data.

Expression Module Table 6.1 shows the datasets that were used for each species.

Table 6.1: Expression datasets considered by the Expression module for the respective species. E-GEOD-3076 is a transcription profiling experiment that profiled the effect of transcription inhibition over a 1 hour time period. E-GEOD-2180 (Baugh et al., 2005) is a transcription profile for 4 different genotypes. E-GEOD-7763 (Chintapalli et al., 2007) is a transcription profile for 8 distinct tissue types (both male and female) and 2 larval tissue types.

Species	**Array Dataset**	**Probe Set**	**Number of Genes**	**Number of Assays**
Yeast	E-GEOD-3076	A-AFFY-27	5943	96
Worm	E-GEOD-2180	A-AFFY-60	11616	123
Fly	E-GEOD-7763	A-AFFY-35	11671	44

Combined Module Table 6.2 shows the number annotations that are considered by the Combined module.

Table 6.2: Number of annotated proteins broken down by annotation type for the respective species. PTMs = Post Translation Modifications.

Species	Pfam	InterPro Motifs	GO Terms	PTMs
Human	22641	17835	22093	6254
Yeast	4890	5143	5843	2599
Worm	11694	15581	10267	244
Fly	16002	20550	9897	712

Orthology All orthologous relationships were downloaded from the InParanoid database (Berglund et al., 2008) and were uploaded into the PIPs 2 database. These orthologues were used for the testing and training of the PIPs predictor for all species considered. The known interactions had already been loaded into the database as part of the Human PIPs predictor work. The number of interactions is shown in Table 6.3. The interactions also include homodimers, which are not considered by the PIPs predictor.

Table 6.3: Experimentally identified protein-protein interactions in different species. Both HPRD and IntAct use high and low throughput data to infer interactions. The interactions present in DIP are based on low throughput experimental data and are of high quality. IntAct does infer interactions for high throughput experiments, such as TAP-TAG which uses spoke expansion for extracting complexes of proteins and inferring the interaction network of the complex, although these can now be filtered out.

Source	**Human**	**Fly**	**Worm**	**Yeast**
DIP	1215	19752	3646	17506
HPRD	33309	—	—	—
IntAct	17374	18523	3468	45565

Even though there are large numbers of interactions for the organisms considered (Table 6.3), the actual overlaps between the databases still remain low (Table 6.4 and Table 6.5).

Table 6.4: Human protein-protein interaction database overlaps.

Source	**DIP**	**HPRD**	**IntAct**
DIP	—	873	170
HPRD		—	3483
IntAct			—

Table 6.5: Species overlap between IntAct and DIP.

Species	**Overlap between DIP and IntAct**
Yeast	7096
Worm	986
Fly	5914

Clustering and Transitive Modules In yeast, as in human, the cluster module was predictive at a likelihood ratio threshold of 5. However, this was not the case for worm and fly models which obtained reduced overall likelihood ratio values. To address the discrepancy, lower likelihood ratio thresholds ranging between 1 to 5 were considered for clustering to identify the optimal threshold for each species.

A similar situation exists for the Transitive modules. Having a likelihood ratio threshold cut off point of 10 results in very few/none of the protein pairs being considered for training or testing. As a result several reduced likelihood ratio thresholds were tried to find an optimal value for the worm and fly.

Modifications to Modules

Modifications were made to each module so that it could accept a taxonomic identifier and then generate the required species-specific predictor. This allowed for the predictor to be called and trained using a command line argument to specify the module to be trained, the training ratio (positive:negative protein pairs) and the species. These modifications provided much easier training and testing of the predictor for each species.

Other than changes to allow for the analysis of different species there were no modifications to algorithms applied by each of the modules (see Chapter 2) and how they calculated likelihood ratios or made predictions. If the predictor was to be extended further and include more species, there would need to be alterations to the Orthology module to handle the increase, in which case it would be worth rewriting sections of the module to be able to handle this.

6.2.2 Results

Calculation of the prior odds ratio was done as described in Chapter 3.2.2. Table 6.6 details the number of positive and negative interactions for each species and the calculated O_{prior}.

Table 6.6: Calculation of the Prior Odds Ratio (O_{prior}). For each species the table shows the number of protein-protein interaction in the positive training dataset and the number of proteins that are annotated as being part of the positive training set. The fourth column is the theoretical maximum number of non-interacting protein pairs (this does not include homo-dimers). The fifth column is the calculated O_{prior}.

Organism	**Positive Interactions**	**Number of Proteins**	**Negative Interactions**	O_{prior}
Yeast	45565	5535	15269780	$\frac{1}{335}$
Worm	3468	2206	2428647	$\frac{1}{700}$
Fly	18523	6695	44797807	$\frac{1}{2418}$

Figure 6.1 shows the predictive capabilities of constructing the PIPs framework within other organisms. Figure 6.1 shows that each of the modules (except Expression) is capable of making predictions above the random. The most predictive of the Expression, Orthology and Combined methods is the Orthologue module based on a ROC100 plot, but in all three species, the Combined module is more predictive over the first 1000 false positive predictions. In contrast, Expression was the weakest of the predictors, with the Yeast expression set failing to make any predictions within the top 100 false positive predictions. The Combined module does prove to be most successful, yeast obtains the highest number of predictions for the first 100 false positives, followed by fly and worm. The combined module most likely performs better for the yeast as it is a well studied organism and therefore has a larger coverage and high volume of annotation for each of its proteins. The Orthology module, although it is accurate, is limited by the number of actual orthologuous and paralogous interactions that are available.

The PIPs predictor was originally designed with human protein-protein interactions in mind, therefore the most accurate modules are within human (see Chapter 3, Figure 3.3, ROC100 scores for the EOCT and EOCM both > 300). The second

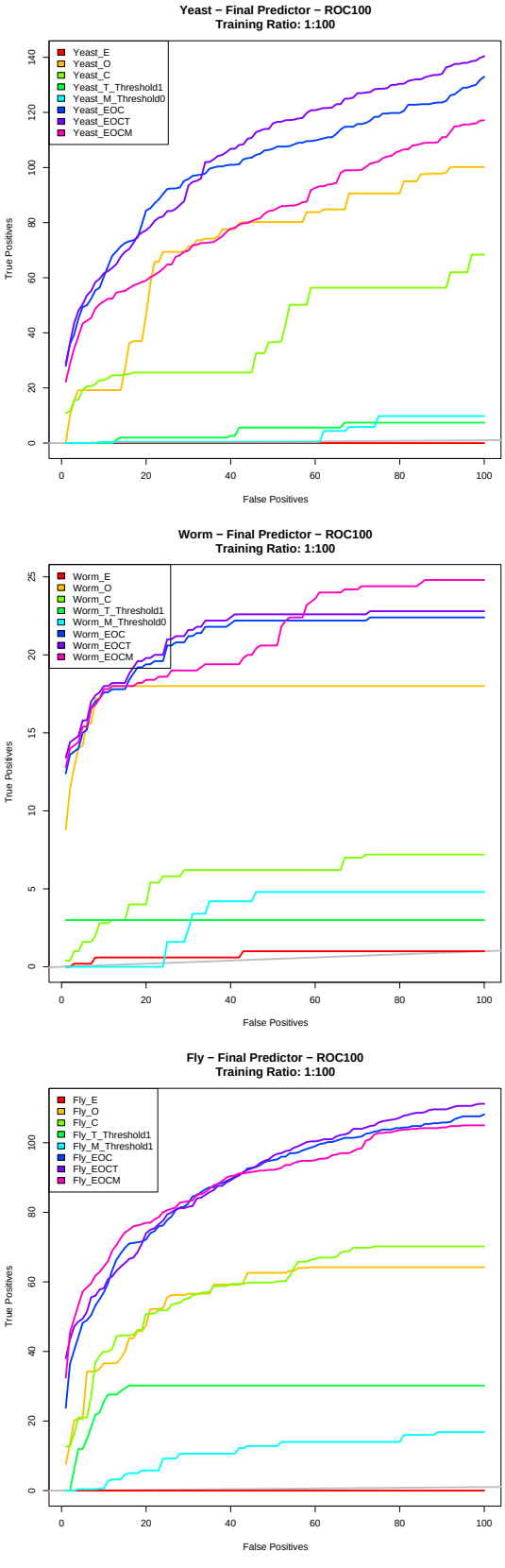


Figure 6.1: ROC100 plots for the final predictors from Yeast, Worm and Fly. Each module in the legends is represented by its single letter code: E = Expression; O = Orthology; C = Combined; T = Transitive; M = Clustering. Where more than one module is involved in the predictions, multiple letters are used to represent the modules that contributed towards the predictions. The Threshold refers to the likelihood ratio threshold used to generate the interaction network used by the Network modules (T and M) based on the predictions from the E, O and C modules. The grey line indicates the performance based on random classification.

most accurate model is the Yeast predictor, likely due to the large amount of research was performed within yeast and the large number of annotations available. Fly and worm are less predictive, but still perform better than random.

6.3 Development of PIPs 2 For Cross Organism Protein-Protein Interaction Prediction

This section describes the development of the PIPs 2 predictor so that is can be trained based on one organism and then tested on the experimental and annotative evidence in a second organism. To test its cross organism predictive capability the PIPs predictor was trained on data from human and then tested on the secondary organism (yeast, fly and worm).

6.3.1 Methods and Data

The Expression and Combined modules used the same datasets as described in Section 6.2.1 for yeast, fly and worm and the same datasets as in Section 3.2.3 for human.

Orthology Module Adaptations

The only module that required modifications to the calculation of the likelihood ratios was the Orthology module. The Orthology module that was used for predictions within the PIPs framework (see Section 2.2.5) had a bin for each species, including a paralogue bin, a bin for occurrence in more than one species and a final bin to allocate pairs that have orthologues but for which there is no experimental evidence to infer that the two proteins interact. For the cross species predictor bins

should not be specified for each species independently. Instead the bins the have been restructured like so:

1. Has orthologue/paralogue, but no known interaction,
2. Has one orthologues/paralogue that has been annotated as interacting,
3. Has two orthologues/paralogues that have been annotated as interacting,
4. Has three orthologues/paralogues that have been annotated as interacting,
5. Has four orthologues/paralogues that have been annotated as interacting.

All other protein pairs were assigned a score of 1.0.

Network Analysis Modules

The code for both the Transitive and Clustering modules remained unchanged. However, the input data likelihood ratios for the training and test protein pairs had to be precalculated. This was done by training the Expression, Orthology and Combined modules and then predicting the training and test dataset values. These scored values were then used to generate the scored datasets that are required by the Transitive and Clustering modules.

Likelihood ratio thresholds were also investigated to determine the protein pairs that are considered for generating a training network for the Transitive and Clustering modules. A range of threshold points were considered to determine the threshold that would be the most predictive based on a ROC100 plot. The new thresholds were then used for testing the modules together. For the Transitive module the thresholds were tested at 2.5, 5 and 10 and for the Clustering module they were

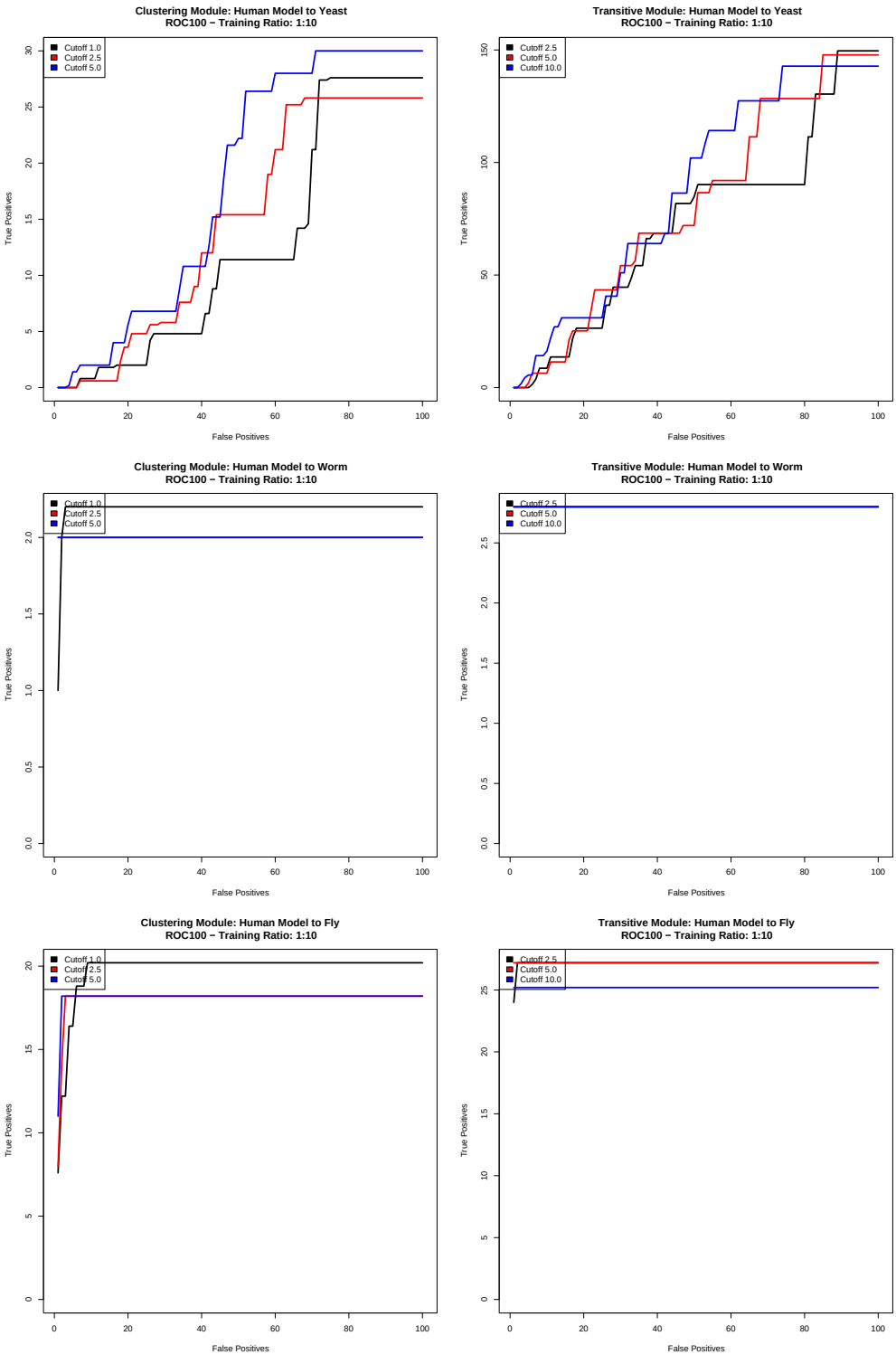
tested at likelihood ratios greater than or equal to 5, 2.5 and 1. Once the optimal thresholds for each species had been selected, these were used for the final testing of the predictor.

6.3.2 Results

Figure 6.2 shows the ROC100 plots for the Clustering and Transitive modules when trained using a variety of likelihood ratio thresholds. For both Fly and Worm the Clustering and Transitive modules do not perform as well as when they are trained in human (see Chapter 4) or yeast. The reason for the poor predictive capability within Fly and worm is most likely due to the poor coverage of the complete proteome for the number of proteins that are known to interact. Based on UniProt figures, the size of the complete proteome for Fly is $\geq$ 19k proteins, and for Worm there are $\geq$ 23k proteins in comparison to the 6695 and 2206 proteins respectively, that are annotated to be interacting. In comparison to yeast where there are $\geq$ 45k interactions involving 5535 proteins and given that there are 5883 proteins within the proteome.

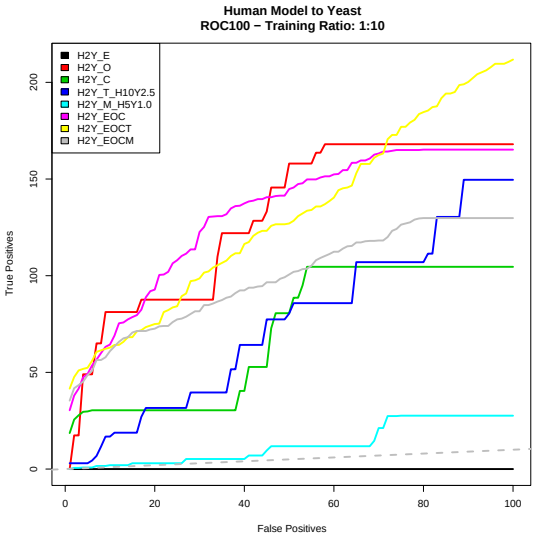
Based on the graphs in Figure 6.2, the threshold values that were selected was a Likelihood Ratio $\geq$ 1.0 for the Clustering Module and $\geq$ 2.5 for the Transitive Module. These thresholds were selected as there is little variation in accuracy of the predictors when changing the threshold. The lowest values were selected as it would not be possible to make such judgements in unknown species, so having an idea of the accuracy given a fixed baseline threshold is more informative. Yeast has the same threshold values as the human predictor, although the Human to Fly and Human to Worm likelihood ratio threshold levels were reduced. Even though the

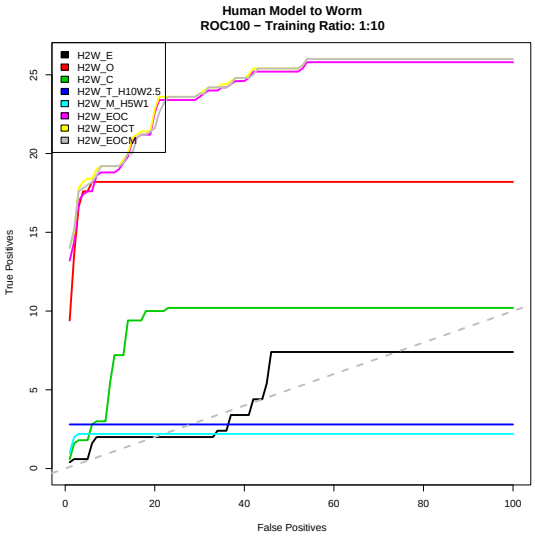
Figure 6.2: ROC100 plots; Left are plots for variation in the likelihood ratio threshold used by the Clustering Module (M); Right are plots for variation in the likelihood ratio threshold used by the Transitive Module (T). From top to bottom by row are the plots for Human to Yeast, Worm and Fly.



Human to Fly and Human to Worm thresholds were reduced, this had little effect on the overall performance of the Clustering and Transitive modules.

The ROC100 plots for training in Human and then testing in Yeast or Fly (Figure 6.3 top and bottom respectively) show that it is possible to predict above random using the final predictors EOCT and EOCM in each case. In Yeast the Transitive and Clustering modules initially affect the predictive capability of the final predictor by lowering the line below that of the EOC predictor. The effect of adding in the networking modules is most likely the result of poor coverage of the proteome during training due to the lack of diversity of the proteins in the positive training set. Even so, it is still striking the effect that this has on the final accuracy of the predictor when the Network modules are included in the predictions.





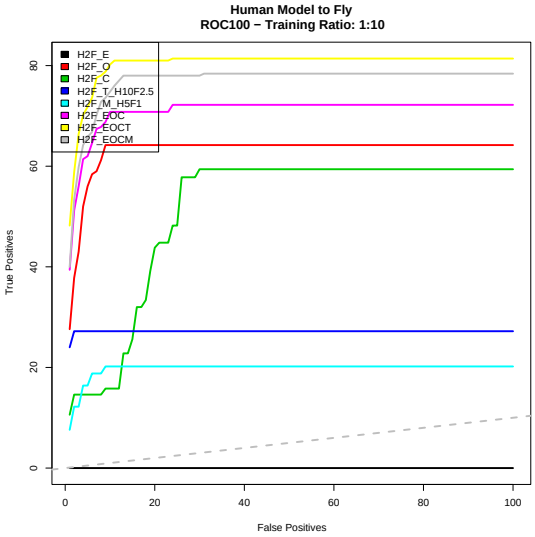


Figure 6.3: ROC100 plots for the final predictors from Human to Yeast, Worm and Fly. Each module in the legends is represented by its single letter code: E = Expression; O = Orthology; C = Combined; T = Transitive; M = Clustering. Where more than one module is involved in the predictions, multiple letters are used to represent the modules that contributed towards the predictions. The dashed grey line indicates the performance based on random classification

6.4 Conclusion

This chapter has shown that the PIPs framework can be applied to different organisms on the provision that there are annotations and experimental datasets available for each of the modules.

The Cross-Species Model Prediction section shows that it is possible to build a PIPs model in one organism that can be applied to another organism with suitable evidence to predict protein-protein interactions. This affords a great opportunity to make predictions in organisms where there are no publically available databases of protein-protein interactions, such as *Trypanosomas*, *Leishmania* and *Xenopus*.

Unlike others (Martin et al., 2005; Shaughnessy et al., 2008), applying the PIPs framework to cross organism protein-protein interaction prediction is relatively successful with Figure 6.3 showing that the final predictor performs better than random. Further development is required to optimise the results and improve the performance of the predictor, especially with regards to the network modules. There is no reason that the PIPs framework could not be applied to organisms that have few publicly annotated protein-protein interactions. The results highlight that there is a difficulty in assigning thresholds for the Transitive and Clustering module. The results can only act as a guide, or ranking system, for the most likely interactions as determining thresholds for the prior probability is problematic due to the lack of information and the ability to estimate the size of an interactome in a targeted organism.

Chapter 7

Jpred Accuracy

Preface

This Chapter investigates the accuracy of the Jpred predictor and analyses whether it is possible to estimate the accuracy of the predictions.

7.1 Introduction

Jpred is a protein secondary structure predictor maintained and developed by the Barton group (Cole et al., 2008). Jpred has been used as part of this study to generate secondary structures for proteins of the human proteome for the development of a sequence module for the PIPs framework (see Section 2.2.4).

The first version of the Jpred predictor (Cuff et al., 1998) used a consensus method to aggregate the predictions of 6 secondary structure predictors. The predictors included: DSC (King and Sternberg, 1996); PHD (Rost and Sander, 1993); NNSSP (Salamov and Solovyev, 1995); PREDATOR (Frishman and Argos, 1997); ZPRED (Zvelebil et al., 1987) and MULPRED (Barton, 1988, unpublished).

The Jpred server was then changed to use the JNet prediction method resulting in

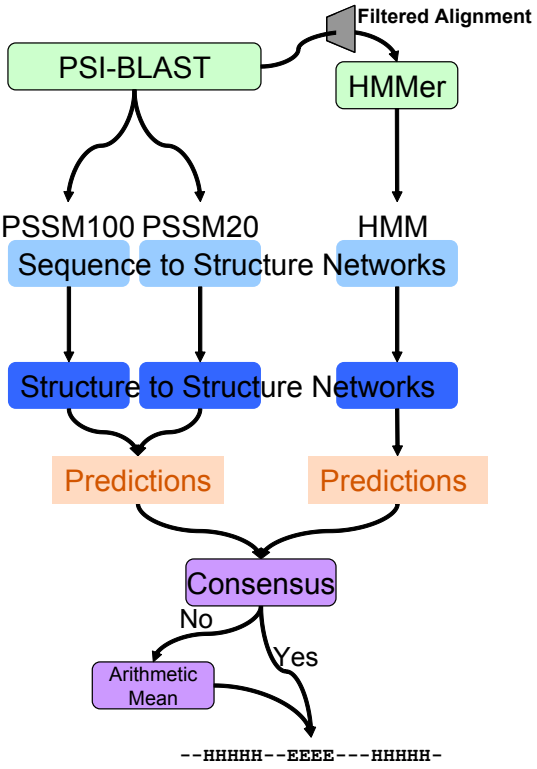


Figure 7.1: A schematic of the JNet 2 Artificial Neural Network architecture. The blue boxes represent individual neural networks and the arrows indicate the flow flow of predictions within the schematic. The inputs to the Sequence to Structure neural networks are based on alignment profiles: PSSMs are Position Specific Scoring Matrices where the suffix refers to the number of hidden nodes in the neural network; and HMM is a Hidden Markov Model from the HMMer package. Two predictions, each consisting of 3 values, one for each secondary structural feature, are calculated and compared to. If the predictions agree this is taken as the final prediction, otherwise the arithmetic mean is calculated and that is given as the final prediction.

an increase in the accuracy of the prediction from 74.6% to 76.4% (Cuff and Barton, 2000). The structural features that Jpred predicts include; α-helical; β-sheet; not αhelical or β-sheet. Improvements have since been made to the predictor to reduce the complexity of the predictor and increase the accuracy of the predictions to 81.5% Cole et al. (2008). Figure 7.1 shows how JNet 2 uses 3 sets of neural networks. JNet 2, as described by Cole et al. (2008), works by querying a sequence with PSI-BLAST against the SwissProt database to generate a multiple sequence alignment (MSA) and a Position Specific Scoring Matrix (PSSM) output. HMMer profiles were formed from the MSA, which was filtered at a 75% sequence similarity cut-off. There are 3 sets of neural networks, two sets use PSSMs, but have varying numbers of hidden nodes within the neural network (PSSM100 has 100 hidden nodes and PSSM20 has 20 hidden nodes) and a third set of neural networks that use the HMMer profiles.

The predictions were made over windowed regions of the protein primary sequence. The two sets of predictions made by each of the neural networks (aggregated PSSM set and the HMM set), each includes 3 values that correspond to the probability of a residue being part of a given secondary structural feature. If all of the outputs agreed for each amino acid then the prediction was fixed, if not the mean of the prediction scores was used to decide the final structural feature of an amino acid.

Jpred provides a reliability/Quality Score for a given prediction for each residue. The Quality score is a measure of the confidence of a structural feature prediction for each residue based on the probabilities assigned by the HMMer neural network predictions (Equation 7.1.1). Figure 7.2 shows the plot of the average accuracy and the average coverage of residues of a blind test set against the Quality score for the JNet predictor (Cuff and Barton, 2000). However, being able to predict the accuracy of a prediction without knowledge of the actual structure of the protein would provide a vital metric to judge the quality of the prediction.

$$QualityScore = Integer(10 \times (out_{max} - out_{next}))$$

Equation 7.1.1: Quality Score for a Jpred prediction, where out_{max} is the score of the highest state and out_{next} is the score of the next highest scoring state calculated by the neural networks (Cuff and Barton, 2000).

This Chapter aims to assess the accuracy of the Jpred predictor and identify whether the quality scores that are provided as part of the predictions can be used as a reliable indicator of the accuracy for the final overall prediction for the secondary structure of the protein or if an alternative method provides a more informative measure.

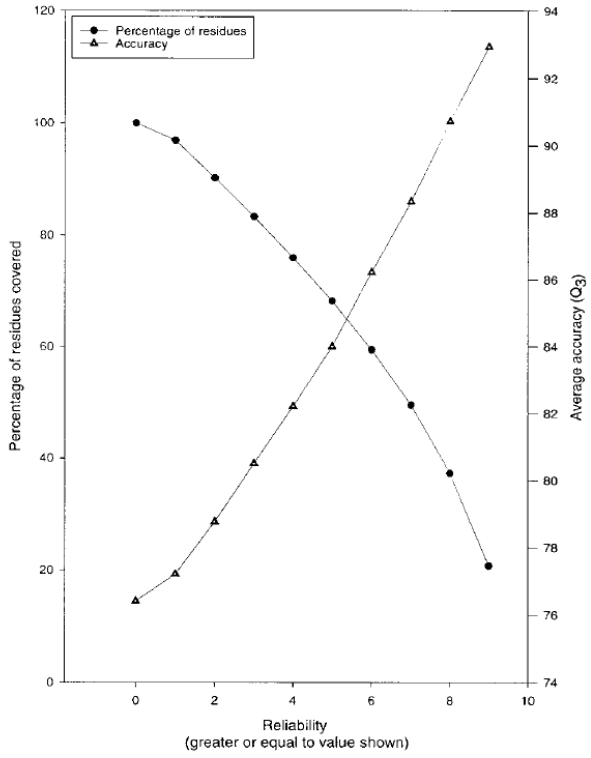


Figure 7.2: The average accuracy (Q_3) and the coverage of residues (%) of the blind test set against the reliability score (Quality Score) from JNet. The diagram is adapted from Cuff and Barton (2000).

7.2 Methods

The predictions that have been used are derived from the new release of the Jpred predictor (Cole et al., 2008). Based on the 3 values predicted by the PSSM and the 3 values by the HMM neural networks to predicted the probability of each structural feature, these are passed to the consensus module, see Figure 7.1. Each residue was binned into one of two 6 dimensional matrices dependent on whether the final prediction was correct. From this it was possible to calculate the probability of any set of Jpred predictions for a predicted protein structure for a given residue being correct. This matrix was then used to score each of the proteins by calculating the mean of the probabilities for each residue.

To avoid over fitting of the data, the Jpred training datasets were used, which allowed for 5 fold cross validation for predicting the accuracy of the assigned secondary structure and determining the robustness of generating an accuracy for a

Jpred prediction. However, due to some data being left out for testing, this did not provide enough coverage across all of the elements in the matrix. To account for missing data, imputation of values was used to fill elements that have missing values. Data was imputed using Nearest Neighbour Hot Decking (Little and Rubin, 1987).

7.2.1 Datasets

Initial analysis was performed on a blind test set used to analyse the performance of the Jpred predictor, consisting of 150 proteins. However, due to the small size of the blind set, further analysis of the data and calculation of a score for the accuracy of a Jpred prediction was performed using the dataset that was used to train the Jpred predictor. The training set contained 1300 known protein secondary structures and Jpred predictions. The training set consisted of 209,940 residues of which 174,196 secondary structural features were correctly predicted by Jpred.

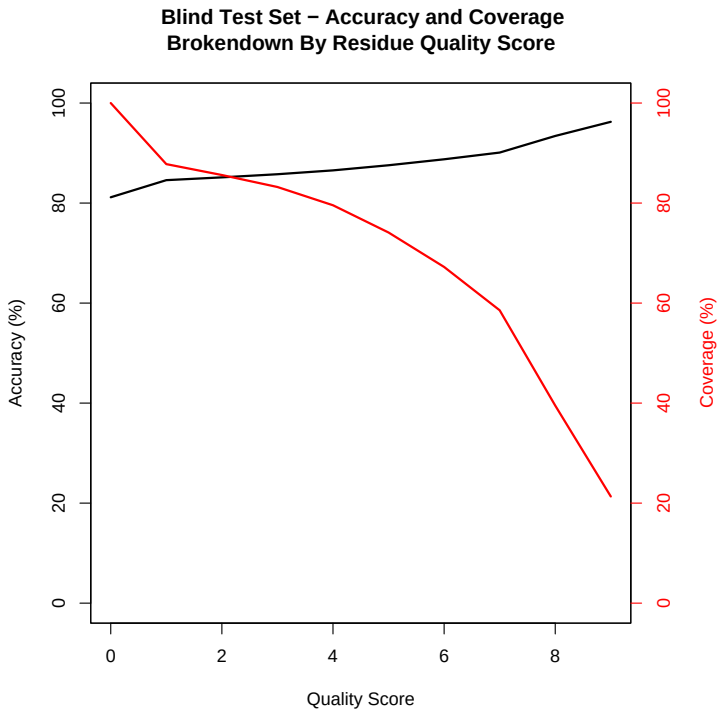
7.2.2 Data Fitting

The *lm* function in R was used to derive a line of best fit between the actual accuracy of the prediction versus the calculated accuracy. The reduced Chi square score was used to determine how well the line of best fit matched the data. A Pearson's correlation coefficient was also calculated to determine how well the calculated accuracy matched the actual accuracy.

7.3 Results

Figure 7.3 indicates the accuracy across all proteins and coverage of the residue space dependent on the set threshold of the quality score for the blind test set. For complete coverage of all residues the mean accuracy is less than 83%, this compares favourably to the blind test set with a mean accuracy of 81%. By increasing the quality score threshold there is a rise in the accuracy (residues with a Quality score of 9 have an accuracy of 97%), however there is a drop in the coverage to 21% of the residues.

Figure 7.3: The black line indicates the accuracy of assigned features per residue dependent on their quality score for the blind test set. The red line indicates the proportion coverage of residues with an assigned quality score for the blind test set.



7.3.1 Average Quality Score

Figure 7.4 shows the average Quality Score for a protein against the actual accuracy of the secondary structure prediction measured using the Q_3 score, where the Q_3 score is the proportion of correctly predicted structural features. The correlation between the accuracy and the average Quality Score is 0.64 (Pearson's).

Figure 7.4: Average Quality Score for a protein plotted against the Accuracy of the prediction as calculated with a Q_3 score for proteins within the training set.

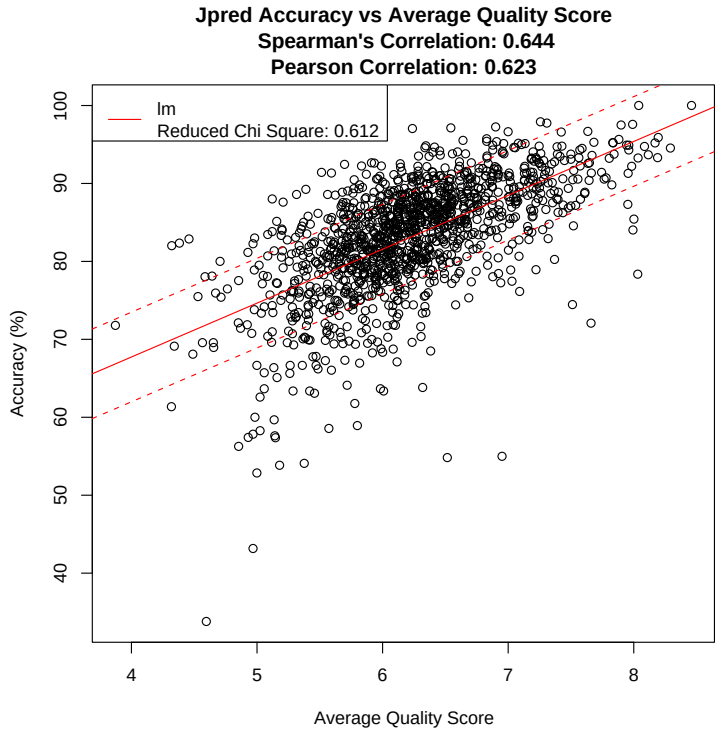
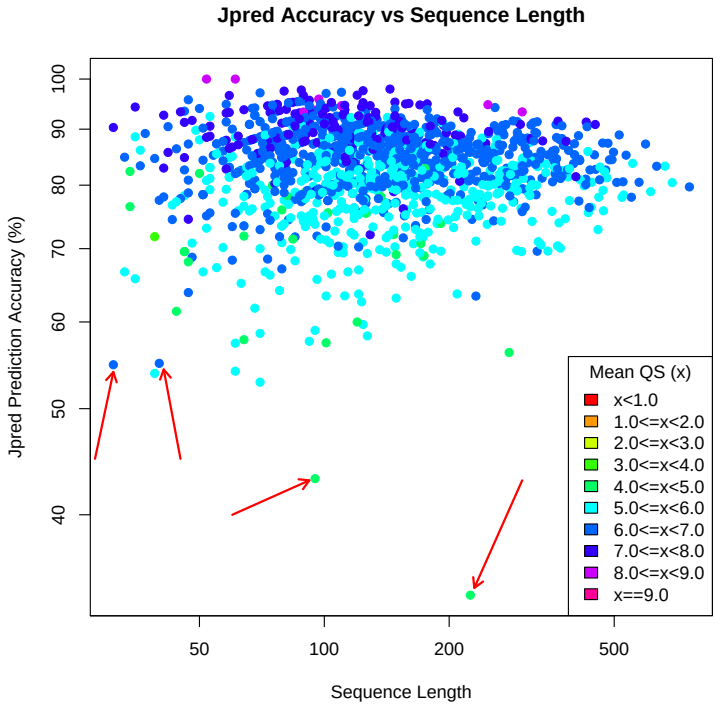


Figure 7.5 compares the accuracy of the prediction against the length of the sequence. There is no correlation between sequence length and the accuracy of the Jpred prediction or of the average quality score assigned by Jpred, Pearson's correlations of -0.02 and -0.05 respectively. The two green points highlighted by the red arrows relate to the proteins Monellin (DSSP: d2o9ux1; Length: 95 residues; Accuracy: 43.16%; Mean Quality Score: 4.97) and OPCA Adhession (DSSP: d2vdfa1;

Figure 7.5: Plot of the protein sequence length against the accuracy of the prediction. The colour indicates the assigned average Quality Score for the prediction of the protein. The arrows highlight proteins of interest.



Length: 225 residues; Accuracy: 33.78; Mean Quality Score: 4.60). These are two proteins Jpred found particularly hard to predict and as such have mean Quality Scores significantly lower than average. A possible explanation for the low quality prediction could be due to the lack of hits to UniRef90 for generating an alignment (Cole et al., 2008).

The two mid blue points in Figure 7.5 that are also highlighted by red arrows are proteins that have received an average mean Quality Score between 6 and 7, but the real accuracy of the prediction is less than 60%. These points correspond to the proteins P53 polypeptide(L) (DSSP: d1aiea) and Mating pheromone ER-1 polypeptide(L) (DSSP: d2erla), which are both short polypeptides (31 and 40 residues in length, respectively). Due to the short nature of the proteins a misclassification of

the secondary structure is going to have a larger effect on the overall accuracy of the prediction in comparison to proteins that are longer with the same number of misclassifications.

7.3.2 Average Probability

Even though there is a correlation between the accuracy and the mean Quality Score, as shown in Figure 7.4, the following Section investigates whether it is possible to improve on the accuracy that is assigned to a Jpred prediction. Where S_{JNet} is the set of 6 probabilities calculated by the two neural networks for each residue it is possible to calculate a likelihood for a residue being assigned a correct secondary structural feature based on a 6 dimensional probability matrix. For each residue set, S_{JNet}, each value is discretised into a number of potential bins, this set of 6 assigned values acts as the coordinate within the 6 dimensional matrix. The mean probability for the correct assignment over all residues of a protein is taken as the average probability for the correct assignment of the secondary structure for the protein. Figure 7.6 shows the effect of using different numbers of bins to calculate a probability. Figure 7.6 shows that there is an increase in the correlation, from 0.397 (Figure 7.6C, 3 bins per dimension) to 0.597 (Figure 7.6A, 10 bins per dimension), between the average assigned probability and the actual accuracy of the secondary structure prediction. However there is an increase in the reduced Chi square score from 0.6 to 0.8 for a fitted line for a reduction in the number of bins, this would suggest that the larger the number of potential bins, then the greater the chance of over fitting the data when calculating a linear regression. This increase in the reduced Chi square score is in part due to the increase in the standard deviation of

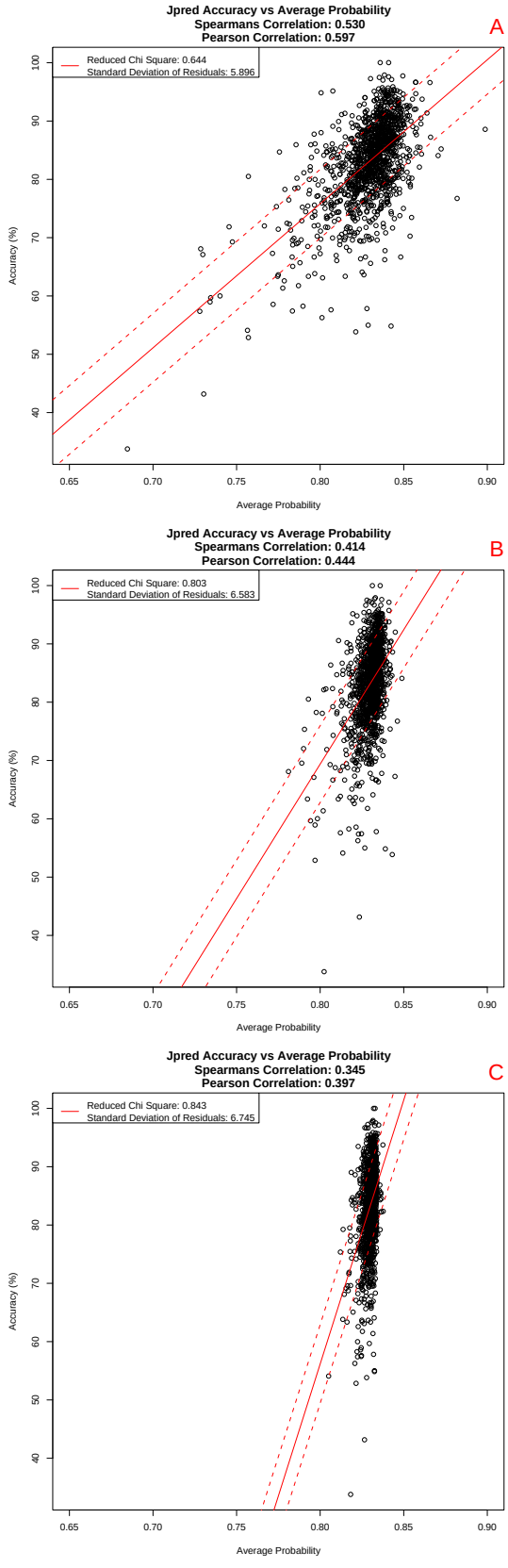
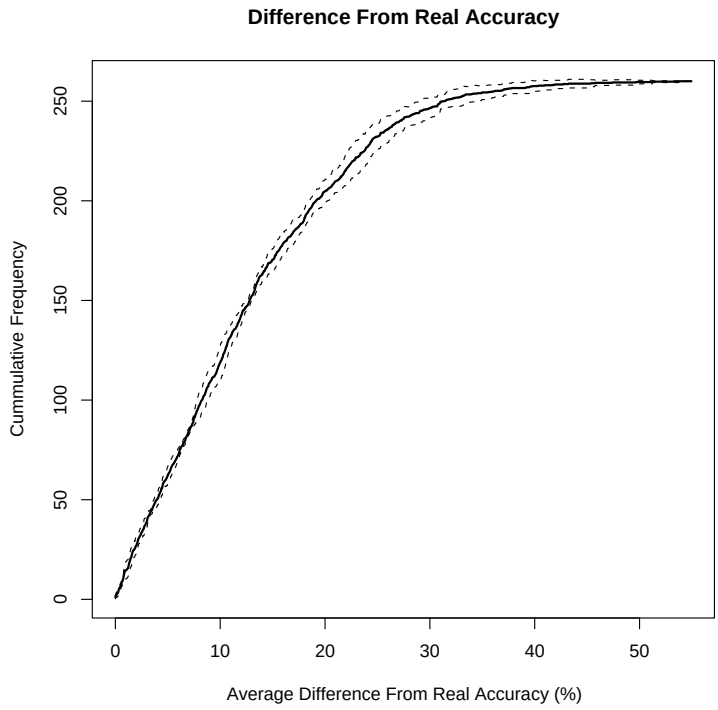


Figure 7.6: Plot of the average probability of the prediction being correct against the actual accuracy of the prediction. A: For each dimension there are 10 bins; B: For each dimension there are 5 bins; C: For each dimension there are 3 bins.

the residuals of the fitted line from 5.9 to 6.8. For further analysis, the number of bins is fixed at 10 bins per dimension as it had a greater correlation with the actual accuracy of the Jpred predictions, even though the reduced Chi square fit suggests that the linear regression is over fitting the observed data.

Using the fitted straight line from Figure 7.6A, accuracies were predicted by calculating the predicted probability that for each residue the assignment of the secondary structure was correct and then calculating the average accuracy for each protein. Figure 7.7 shows the deviation of the predicted accuracy from the real accuracy during 5 fold cross validation. The plot shows that 68% of the predictions generate an accuracy that has an error of ±15.8% and 95% of the predictions have accuracy with an error of ±30.4%. The mean difference of the predicted accuracy

Figure 7.7: Deviation of predicted accuracy from the real accuracy using 5 fold cross validation. The dotted line show 1 standard deviation from the mean.



to the real accuracy is ±12.7%.

Including the residue quality score along with the 6 feature probabilities in S_{JNet}, in a 7 dimensional matrix increases the correlation between the calculated probability of a correct prediction and the actual accuracy of the prediction, see Figure 7.8. The Pearson's correlation increases to 0.79 in comparison to 0.64 and 0.59 (mean probability (neural network inputs only) and mean quality score alone respectively). However, this increase in the correlation is met with a decrease in the reduced Chi square for a fitted line (0.39). The decrease in the reduced Chi square score suggests that the predicted values are over fitting the actual accuracy

By discretising the predicted accuracy (Figure 7.8 and Table 7.1) in relation to the real accuracy it is possible to get a more accurate measure on the error associated to the predicted accuracy of the prediction, this shows that assigned probabilities correspond well to the actual accuracies above 0.7. Figure 7.9 and Table 7.2 show that it is also possible to use the mean quality score to predict the accuracy of a secondary structure prediction. Using the mean quality score does not handle predictions that have low accuracy ($< 60\%$), it would therefore result in the over estimation of the accuracy for some proteins. Both the average quality score and average probability are less capable at making predictions of accuracy for low quality predictions, this is due to the lack of examples that have lower accuracy predictions.

Figure 7.8: Average probability, including Quality Score (7 dimensional datset), with discretised accuracies based on the calculated probability of the Jpred prediction being correct.

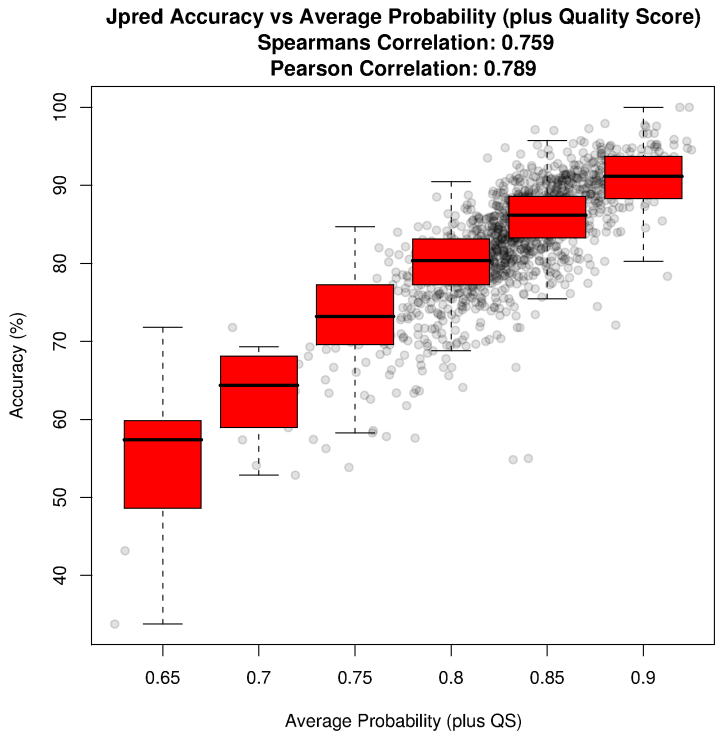


Table 7.1: Relationship of predicted accuracies, based on the 7 dimensional dataset, within a given range to the mean, standard deviation and 1^{st} and 3^{rd} quartiles for the real accuracies of the secondary structure predictions.

Predicted Accuracy (x)	Real Accuracy			
	Mean	Standard Deviation	1st Quartile	3rd Quartile
$x < 0.675$	54.27	12.4	48.63	59.84
$0.675 \leq x < 0.725$	63.44	5.3	59.55	67.84
$0.725 \leq x < 0.775$	72.69	6.4	69.57	77.22
$0.775 \leq x < 0.825$	80.02	4.8	77.27	83.12
$0.825 \leq x < 0.875$	85.7	4.6	83.25	88.60
$0.875 \leq x$	90.67	4.3	88.30	93.72

Figure 7.9: Average Quality Score with discretised accuracies based on the calculated probability of the Jpred prediction being correct.

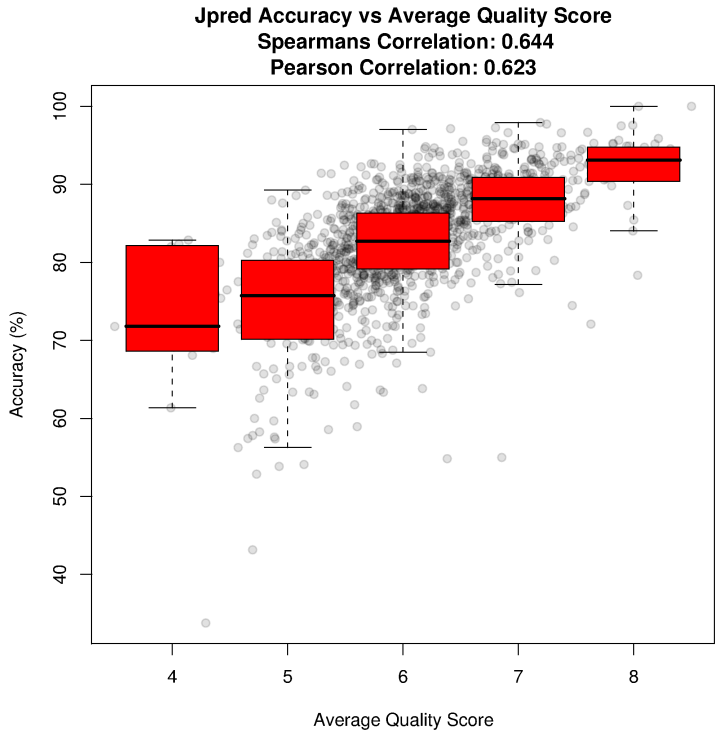


Table 7.2: Relationship of average quality score within a given range to the mean, standard deviation and 1^{st} and 3^{rd} quartiles for the real accuracies of the secondary structure predictions.

Average Quality Score (x)	Real Accuracy			
	Mean	Standard Deviation	1st Quartile	3rd Quartile
$x < 4.5$	73.94	8.52	68.61	82.18
$4.5 \leq x < 5.5$	74.23	8.74	70.16	80.13
$5.5 \leq x < 6.5$	82.35	5.68	79.15	86.30
$6.5 \leq x < 7.5$	87.72	5.10	85.28	90.91
$7.5 \leq x < 8.5$	91.54	5.39	90.38	94.76
$8.5 \leq x$	—	—	—	—

7.4 Conclusion

This Chapter shows that it is possible to predict the accuracy of a secondary structure prediction that has been made by Jpred based on the mean probabilities from the two neural networks. This is important as it is the first time that this has been possible and allows for an accurate calculation of the quality of the final prediction.

One of the potential problems with this work is that the accuracy of the predictions has been measured using the Q_3 measure of accuracy (Equation 7.4.1).

$$Q_3 = \frac{w_i}{l}$$

Equation 7.4.1: Q_3 measure of accuracy where w is the number of residues correctly assigned secondary structure predictions when i is α, β or not α or β and l is the length of the protein.

The Q_3 measure of accuracy has been shown to over estimate the actual accuracy of a secondary structure prediction (Zhang and Zhang, 2001, 2003). Other methods have been suggested, such as Q_9 that account for the proportions of each feature that exist within the protein. Further work upon this method is required to also account for changing the measure of accuracy from the Q_3 to Q_9 measure.

The second limitation with the current method is that the predictions of accuracy have been made based on the training set, as such the actual accuracies are likely to be overestimated. This was due to the limited number of blind set test structures, but it is the trend that is important for estimating the accuracy of a prediction that has been made by Jpred.

Chapter 8

Discussion and Conclusions

Preface

This Chapter sumarises the major developments achieved during the thesis and proposes future directions for the project.

8.1 The PIPs Framework

8.1.1 Modules

Combined Module

The Combined module provides the scope to integrate new information based on diverse annotative forms of evidence and is capable of predicting a high likelihood of interaction between protein pairs, both in humans (see Section 2.2.1) and in other species (see Section 6.2). The Combined module has also proved to be an effective cross species predictor (see Section 6.3). Chapter 2.2.1 shows the removal of the co-localisation information from the Combined module does not degrade the quality of the predictions that are made. However, this allows for protein localisation information to be used to analyse the predictions that have been made, for example

Figure 4.17 indicates that for the predictions that were made, the protein pairs tended to co-localise within the cell.

Network Modules

The development of the Clustering module has resulted in an increase in the number of true positive predictions over the first 100 false positive predictions that are made by the PIPs 2 predictor with the Clustering module in comparison to PIPs 1 with the Transitive module. Figure 2.10 highlights the improved accuracy over the PIPs 1. Both the Transitive and Clustering modules are predictive independent of each other, but there is a correlation in the predictions that are made (Pearson correlation coefficient of 0.25). Therefore the two modules could not be used in conjunction as part of the PIPs 2 predictor. One way around this would be to incorporate the predictions of both the Clustering and Transitive modules into a single Network Module within a full Bayesian construct. This negates issues of correlation of predictions made by the two networking analysis modules and the evidence from both could be included within PIPs predictor. This would also mean that a single final score for the likelihood of interaction between protein pairs would be calculated. Having a single score would also be beneficial for the end user as they do not have to make a judgement about which score to use. The modules were not integrated into a single module due to a time constraints.

Expression Module

Section 2.2.3 shows that the gene-expression data can be used for the prediction of protein-protein interaction, but the correct selection of experimental datasets is crucial for making predictions. Future work should focus on finding larger datasets

that have greater coverage of the proteome or moving towards using Next Generation Sequencing data where the reads represent what is actually being transcribed within the cell at any one time. One advantage for using next generation sequencing data is that there is a reading for everything that is transcribed within the cell and not a preselected number of genes that have probes on a chip, which is the limitation with the current microarray technologies.

Sequence Module

Even though the Sequence module (see Section 2.2.4) was not included in the final PIPs predictor due to low accuracy of the predictions, this does not mean that it should not be reinvestigated for inclusion within the predictor at a later time. It is the primary sequence of protein that leads to eventual structure and physical and chemical properties. The primary sequence of a protein therefore holds a lot of information about the properties of the protein and the potential protein interaction interfaces. Section 2.2.4 shows it is a non-trivial task to filter the information and represent it in a meaningful and predictive form for the identification of protein-protein interaction.

Although no further development was done for the Sequence module, there are many further improvements that could be made to increase its predictive capabilities. The current Sequence module considers all residues within a protein so the first step would be to consider only the properties of exposed residues. Residues that are buried deep within the protein would have a lower probability of interacting with the residues of another protein.

8.1.2 Final Predictions

The PIPs 2 predictor has increased the number of predictions made by the PIPs 1 predictor from 37,606 to 310,893 potential protein-protein interactions. Of the predictions that are made by the PIPs 2 predictor, Figures 4.6 and 4.7 show that protein pairs that are predicted to interact have significantly higher likelihood ratios if they are part of the same biological process/pathway, but this would be expected as Gene Ontology terms for Biological Process where included in the Combined module. Co-location information was not included during the prediction process, but Figure 4.17 shows that protein pairs that are predicted to interact also tend to co-locate.

When it came to calculating the final set of protein-protein interaction predictions allowing the predictor to train each module separately had several advantages. The first was that there was resilience to failure during training and predicting, so if one module had a problem causing the training to fail, only that module would be affected and could easily be re-run without having to retrain other modules. The other advantage was to be able to make a full set of predictions for the Expression, Orthology and Combined modules and use that final set of predictions within the Clustering and Transitive network modules.

Modifications were also made to the Orthology to increase the speed of the predictor. For the Orthology module, this was done by moving the orthologous pairs to being held in memory rather than queried from a database. This provides an increase in the predictions as it is quicker to return information from memory than to query a database over a network.

8.2 Cross Organism Protein-Protein Interaction Prediction

The PIPs 2 predictor is able to make predictions in organisms other than human. The PIPs 2 predictor is also capable of making predictions in organisms that have no annotated interactions (see Section 6.3). This makes it possible to predict protein-protein interactions in organisms that have become the focus of recent studies, but do not have annotated interactions.

One downside in using models that have been trained in one organism and applied in a second organism is the difficulty in estimating the prior odds ratio. However, the PIPs 2 predictor does allow for the creation of a ranked list of probable protein-protein interactions based on the available evidence for a new organism. Organisms that would greatly benefit from the use of the PIPs 2 predictor would be the Xenopus, Trypanosoma or Leishmania organisms as there is readily available information about gene expression, Gene Ontology terms, protein domains and post translational modifications and Orthologous interactions.

8.3 Future Work

The future of the PIPs 2 predictor should be to take the predicted interactome and identify the false positive predictions by eliminating interactions based on computational reasoning. This means removing predictions that are not likely to occur based on biological circumstances, such as temporal differences, physical exclusion (e.g. interactions within a complex) or location within a cell.

The predictions that have been made are a snap-shot of all potential interactions

that can occur. One way to get around this is to develop an interactome to model biological processes by iteratively refining the set of interactions until certain criteria are met. Allowing the interactome to remove or add interactions to emulate annotated biological pathways would result in the removal of many of the false positive predictions and act as a starting point to modelling a cell *in silico*.

The end result would be to identify pairs of proteins that are predicted to interact given a set of criteria, such as localisation or the requirement for previous interactions to have occurred. In the future this could help determine the robustness of cells to alterations in the environment and lead to a more targeted approach to drug development.

There is an increase in the rate of genomes that are being sequenced and the number of available high throughput methods for analysis of properties of cells at the proteome level. The identification of protein-protein interactions will become more data driven by analysis of large scale high throughput techniques for the identification of the most likely interactions, which can then be verified. This changes the focus from identifying interactions because they are involved in the same pathway to identifying interactions then determining their biological significance.

For cross species protein-protein interaction prediction future work should investigate using different training organisms for making predictions. The hypothesis is that organisms that are more closely related to a target organism are more likely to share similar predictive models. However, as has been shown with training PIPs based on worm, the limiting factor is a lack of prior knowledge about an organism for generating an accurate predictor. It is important to understand that the training organism and the target organism have evolutionarily diverged from each other.

This divergence may play a key role in the effectiveness of the predictor. There may also be a fine balance between training in an organism that is more closely related to the target organism, but has less prior knowledge in comparison to a further relative that is more heavily annotated.

In summary the thesis describes the improvements in the prediction of protein-protein interactions implemented as part of the naïve Bayesian framework of the PIPs 2 predictor. It also highlights the capability of the predictor to be applied outside the prediction of interactions in a single organism to predict interactions in multiple organisms. While there is a lot of work to be done to refine the predictions, this thesis indicates the potential future direction of protein-protein interaction prediction.

Bibliography

Aebersold, Ruedi, and Matthias Mann. 2003. Mass spectrometry-based proteomics. *Nature* 422(6928):198–207.

Affymetrix. 2002. Statistical algorithms description document.

Alfarano, C, C E Andrade, K Anthony, N Bahroos, M Bajec, K Bantoft, D Betel, B Bobechko, K Boutilier, E Burgess, K Buzadzija, R Cavero, C D'Abreo, I Donaldson, D Dorairajoo, M J Dumontier, M R Dumontier, V Earles, R Farrall, H Feldman, E Garderman, Y Gong, R Gonzaga, V Grytsan, E Gryz, V Gu, E Haldorsen, A Halupa, R Haw, A Hrvojic, L Hurrell, R Isserlin, F Jack, F Juma, A Khan, T Kon, S Konopinsky, V Le, E Lee, S Ling, M Magidin, J Moniakis, J Montojo, S Moore, B Muskat, I Ng, J P Paraiso, B Parker, G Pintilie, R Pirone, J J Salama, S Sgro, T Shan, Y Shu, J Siew, D Skinner, K Snyder, R Stasiuk, D Strumpf, B Tuekam, S Tao, Z Wang, M White, R Willis, C Wolting, S Wong, A Wrong, C Xin, R Yao, B Yates, S Zhang, K Zheng, T Pawson, B F F Ouellette, and C W V Hogue. 2005. The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Research* 33(Database issue):D418–424.

Altschul, S F, T L Madden, A A Schffer, J Zhang, Z Zhang, W Miller, and D J Lipman. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research* 25(17):3389–3402.

Aragues, Ramon, Daniel Jaeggi, and Baldo Oliva. 2006. Piana: protein interactions and network analysis. *Bioinformatics* 22(8):1015–1017.

Aranda, B., P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob. 2009. The intact molecular interaction database in 2010. *Nucl. Acids Res.* gkp878.

Arifuzzaman, Mohammad, Maki Maeda, Aya Itoh, Kensaku Nishikata, Chiharu Takita, Rintaro Saito, Takeshi Ara, Kenji Nakahigashi, Hsuan-Cheng Huang, Aki Hirai, Kohei Tsuzuki, Seira Nakamura, Mohammad Altaf-Ul-Amin, Taku Oshima, Tomoya Baba, Natsuko Yamamoto, Tomoyo Kawamura, Tomoko Ioka-Nakamichi, Masanari Kitagawa, Masaru Tomita, Shigehiko Kanaya, Chieko Wada, and Hirotada Mori. 2006. Large-scale identification of protein-protein interaction of escherichia coli k-12. *Genome Res* 16(5):686–691.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet* 25(1):25–29.

Bader, Gary, and Christopher Hogue. 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4(1):2.

Barabasi, Albert-Laszlo, and Zoltan N. Oltvai. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5(2):101–113.

Barnard, G. A. 1949. Statistical inference. *Journal of the Royal Statistical Society. Series B (Methodological)* 11(2):115–149.

Barrell, Daniel, Emily Dimmer, Rachael P. Huntley, David Binns, Claire O'Donovan, and Rolf Apweiler. 2009. The goa database in 2009–an integrated gene ontology annotation resource. *Nucl. Acids Res.* 37(suppl 1):D396–403.

Barrett, Tanya, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. 2007. Ncbi geo: mining tens of millions of expression profiles–database and tools update. *Nucl. Acids Res.* 35(suppl 1):D760–765.

Bartel, P. L., J. A. Roecklein, D. SenGupta, and S. Fields. 1996. A protein linkage map of escherichia coli bacteriophage t7. *Nat Genet* 12(1):72–77.

Baugh, L. Ryan, Andrew A. Hill, Julia M. Claggett, Kate Hill-Harfe, Joanne C. Wen, Donna K. Slonim, Eugene L. Brown, and Craig P. Hunter. 2005. The homeodomain protein pal-1 specifies a lineage-specific regulatory network in the c. elegans embryo. *Development* 132(8):1843–1854.

Bayes, Thomas, and Richard Price. 1763. An essay towards solving a problem in the doctrine of chances. *Phil* 53:370–418.

Ben-Hur, Asa, and William Stafford Noble. 2005. Kernel methods for predicting protein-protein interactions. *Bioinformatics* 21(suppl 1):i38–46.

———. 2006. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics* 7 Suppl 1:S2.

Berglund, Ann-Charlotte, Erik Sjlund, Gabriel Ostlund, and Erik L L Sonnhammer. 2008. Inparanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res* 36(Database issue):D263–D266.

Bernas, Tytus, Grald Grgori, Eli K Asem, and J. Paul Robinson. 2006. Integrating cytomics and proteomics. *Mol Cell Proteomics* 5(1):2–13.

Bhardwaj, Nitin, and Hui Lu. 2005. Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics* 21(11): 2730–2738.

Blatt, Wiseman, and Domany. 1996. Superparamagnetic clustering of data. *Phys Rev Lett* 76(18):3251–3254.

Blumenthal, T. 1998. Gene clusters and polycistronic transcription in eukaryotes. *Bioessays* 20(6):480–487.

Bolser, Dan, Panos Dafas, Richard Harrington, Jong Park, and Michael Schroeder. 2003. Visualisation and graph-theoretic analysis of a large-scale protein structural interactome. *BMC Bioinformatics* 4(1):45.

Bolstad, B.M., R.A Irizarry, M. Astrand, and T.P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2):185–193.

Bonetta, Laura. 2005. Flow cytometry smaller and better. *Nat Meth* 2(10):785–795.

Bonvin, Alexandre M J J, Rolf Boelens, and Robert Kaptein. 2005. Nmr analysis of protein interactions. *Curr Opin Chem Biol* 9(5):501–508.

Borch, Jonas, Thomas J D Jrgensen, and Peter Roepstorff. 2005. Mass spectrometric analysis of protein interactions. *Curr Opin Chem Biol* 9(5):509–516.

Boser, BE, IM Guyon, and VN Vapnik. 1992. A training algorithm for optimal margin classifiers. *5th Annual ACM Workshop on COLT* 144–152.

Breitkreutz, Bobby-Joe, Chris Stark, and Mike Tyers. 2003. The grid: the general repository for interaction datasets. *Genome Biology* 4(3):R23.

Brohee, Sylvain, and Jacques van Helden. 2006. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 7(1):488.

Brown, Kevin R., and Igor Jurisica. 2005. Online predicted human interaction database. *Bioinformatics* 21(9):2076–2082.

Burger, Lukas, and Erik van Nimwegen. 2008. Accurate prediction of protein-protein interactions from sequence alignments using a bayesian method. *Mol Syst Biol* 4.

Burkhardt, Janis K, Esteban Carrizosa, and Meredith H Shaffer. 2008. The actin cytoskeleton in t cell activation. *Annual Review of Immunology* 26:233–259.

Carrizosa, Esteban, Timothy S. Gomez, Christine M. Labno, Deborah A. Klos Dehring, Xiaohong Liu, Bruce D. Freedman, Daniel D. Billadeau, and Janis K. Burkhardt. 2009. Hematopoietic lineage cell-specific protein 1 is recruited to the immunological synapse by il-2-inducible t cell kinase and regulates phospholipase cgamma1 microcluster dynamics during t cell spreading. *J Immunol* 183(11):7352–7361.

Castillo, Andrea R, Janet B Meehl, Garry Morgan, Amy Schutz-Geschwender, and Mark Winey. 2002. The yeast protein kinase mps1p is required for assembly of the integral spindle pole body component spc42p. *The Journal of Cell Biology* 156(3):453–465.

Ceol, Arnaud, Andrew Chatr Aryamontri, Luana Licata, Daniele Peluso, Leonardo Briganti, Livia Perfetto, Luisa Castagnoli, and Gianni Cesareni. 2010. Mint, the molecular interaction database: 2009 update. *Nucleic Acids Research* 38(Database issue):D532–539.

Chakrabarti, Pinak, and Jol Janin. 2002. Dissecting protein-protein recognition sites. *Proteins* 47(3):334–343.

Chatr-aryamontri, Andrew, Arnaud Ceol, Luisa Montecchi Palazzi, Giuliano Nardelli, Maria Victoria Schneider, Luisa Castagnoli, and Gianni Cesareni. 2007. Mint: the molecular interaction database. *Nucleic Acids Res* 35(Database issue): D572–D574.

Cheng, Y., and G. M. Church. 2000. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* 8:93–103.

Chien, C, PL Bartel, R Sternglanz, and S Fields. 1991. The two-hybrid system: A method to identify and clone genes for proteins that interact with a protein of interest. *Proceedings of the National Academy of Sciences* 88(21):9578–9582.

Chinnasamy, Arunkumar, Ankush Mittal, and Wing-Kin Sung. 2006. Probabilistic prediction of protein-protein interactions from the protein sequences. *Computers in Biology and Medicine* 36(10):1143–1154.

Chintapalli, Venkateswara R, Jing Wang, and Julian A T Dow. 2007. Using flyatlas to identify better drosophila melanogaster models of human disease. *Nat Genet* 39(6):715–720.

Clauset, Aaron, Cristopher Moore, and M. E. J. Newman. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* 453(7191):98–101.

Cochrane, Guy, Ilene Karsch-Mizrachi, Yasukazu Nakamura, and International Nucleotide Sequence Database Collaboration. 2011. The international nucleotide sequence database collaboration. *Nucleic Acids Res* 39(Database issue):D15–D18.

Cole, Christian, Jonathan D. Barber, and Geoffrey J. Barton. 2008. The jpred 3 secondary structure prediction server. *Nucl. Acids Res.* 36(suppl 2):W197–201.

Consortium, UniProt. 2008. The universal protein resource (uniprot). *Nucleic Acids Research* 36(Database issue):D190–195.

Conte, L. Lo, C. Chothia, and J. Janin. 1999. The atomic structure of protein-protein recognition sites. *J Mol Biol* 285(5):2177–2198.

Couto, Francisco M., Mrio J. Silva, and Pedro M. Coutinho. 2007. Measuring semantic similarity between gene ontology terms. *Data & Knowledge Engineering* 61(1):137–152.

Cox, D R. 2006. *Frequentist and bayesian statistics: A critique (keynot address).*

Cuff, J A, and G J Barton. 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 40(3):502–511.

Cuff, J A, M E Clamp, A S Siddiqui, M Finlay, and G J Barton. 1998. Jpred: a consensus secondary structure prediction server. *Bioinformatics (Oxford, England)* 14(10):892–893.

Cunningham, B. C., and J. A. Wells. 1989. High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis. *Science* 244(4908): 1081–1085.

Cusick, Michael E, Haiyuan Yu, Alex Smolyar, Kavitha Venkatesan, Anne-Ruxandra Carvunis, Nicolas Simonis, Jean-Francois Rual, Heather Borick, Pascal Braun, Matija Dreze, Jean Vandenhaute, Mary Galli, Junshi Yazaki, David E Hill, Joseph R Ecker, Frederick P Roth, and Marc Vidal. 2009. Literature-curated protein interaction datasets. *Nat Meth* 6(1):39–46.

Dandekar, T, Y Snel, M Heynen, and P Bork. 1999. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biological Chemistry* 23:324–328.

Date, Shailesh V, and Edward M Marcotte. 2005. Protein function prediction using the protein link explorer (plex). *Bioinformatics* 21(10):2558–2559.

Deng, Lei, Jihong Guan, Qiwen Dong, and Shuigeng Zhou. 2009. Prediction of protein-protein interaction sites using an ensemble method. *BMC Bioinformatics* 10(1):426.

D'haeseleer, Patrik, and George M Church. 2004. Estimating and improving protein interaction error rates. *Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference* 216–223.

Dongen, Stijn Marinus van. 2000. Graph clustering by flow simulation. Ph.D. thesis, University of Utrecht.

Drawid, A, and M Gerstein. 2000. A bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *Journal of Molecular Biology* 301(4):1059–1075.

Enright, A J, I Iliopoulos, N C Kyrpides, and C A Ouzounis. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402(6757): 86–90.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.* 30(7):1575–1584.

Fariselli, Piero, Florencio Pazos, Alfonso Valencia, and Rita Casadio. 2002. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem* 269(5):1356–1361.

Fawcett, Tom. 2006. An introduction to roc analysis. *Pattern Recognition Letters* 27(8):861–874.

Fernandez, Jose M., Robert Hoffmann, and Alfonso Valencia. 2007. ihop web services. *Nucl. Acids Res.* 35(suppl 2):W21–26.

Fields, S., and O. Song. 1989. A novel genetic system to detect protein-protein interactions. *Nature* 340(6230):245–246.

Fields, Stanley. 2009. Interactive learning: Lessons from two hybrids over two decades. *PROTEOMICS* 9(23):52095213.

Finn, Robert D., Mhairi Marshall, and Alex Bateman. 2005. ipfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics* 21(3):410–412.

Finn, Robert D, Jaina Mistry, Benjamin Schuster-Bckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, Sean R Eddy, Erik L L Sonnhammer, and Alex Bateman. 2006. Pfam: clans, web tools and services. *Nucleic Acids Research* 34(Database issue): D247–251.

Fisher, R. A. 1930. Inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society* 26(04):528–535.

Formstecher, Etienne, Sandra Aresta, Vincent Collura, Alexandre Hamburger, Alain Meil, Alexandra Trehin, Cline Reverdy, Virginie Betin, Sophie Maire, Christine Brun, Bernard Jacq, Monique Arpin, Yohanns Bellaiche, Saverio Bellusci, Philippe Benaroch, Michel Bornens, Roland Chanet, Philippe Chavrier, Olivier Delattre, Valrie Doye, Richard Fehon, Grard Faye, Thierry Galli, Jean-Antoine Girault, Bruno Goud, Jean de Gunzburg, Ludger Johannes, Marie-Pierre Junier, Vincent Mirouse, Ashim Mukherjee, Dora Papadopoulo, Franck Perez, Anne Plessis, Carine Ross, Simon Saule, Dominique Stoppa-Lyonnet, Alain Vincent, Michael White, Pierre Legrain, Jrme Wojcik, Jacques Camonis, and Laurent Daviet. 2005. Protein interaction mapping: a drosophila case study. *Genome Res* 15(3):376–384.

Friedman, Nir, Dan Geiger, and Moises Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning* 29(2):131–163.

Frishman, D, and P Argos. 1997. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 27(3):329–335.

Gandhi, T K B, Jun Zhong, Suresh Mathivanan, L Karthick, K N Chandrika, S Sujatha Mohan, Salil Sharma, Stefan Pinkert, Shilpa Nagaraju, Balamurugan Periaswamy, Goparani Mishra, Kannabiran Nandakumar, Beiyi Shen, Nandan Deshpande, Rashmi Nayak, Malabika Sarker, Jef D Boeke, Giovanni Parmigiani, Jorg Schultz, Joel S Bader, and Akhilesh Pandey. 2006. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* 38(3):285–293.

Gentleman, Robert, Vincent Carey, Douglas Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry,

Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Yang, and Jianhua Zhang. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* 5(10):R80.

Gerstein, Mark, Ning Lan, and Ronald Jansen. 2002. Proteomics: Enhanced: Integrating interactomes. *Science* 295(5553):284–287.

Giot, L., J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. 2003. A protein interaction map of drosophila melanogaster. *Science* 302(5651):1727–1736.

Gomez, Shawn M, William Stafford Noble, and Andrey Rzhetsky. 2003. Learning to predict protein-protein interactions from protein sequences. *Bioinformatics* 19(15):1875–1881.

Gu, Jiajun, and Jun S Liu. 2008. Bayesian biclustering of gene expression data. *BMC Genomics* 9 Suppl 1:S4.

Hardin, Johanna, Aya Mitani, Leanne Hicks, and Brian VanKoten. 2007. A robust measure of correlation between two genes on a microarray. *BMC Bioinformatics* 8(1):220.

Hart, G Traver, Arun Ramani, and Edward Marcotte. 2006. How complete are current yeast and human protein-interaction networks? *Genome Biology* 7(11): 120.

He, Min, Yi Wang, and Wei Li. 2009. Ppi finder: A mining tool for human protein-protein interactions. *PLoS ONE* 4(2):e4554.

Hegyi, Hedi, Eva Schad, and Peter Tompa. 2007. Structural disorder promotes assembly of protein complexes. *BMC Struct Biol* 7:65.

Huang, Tao-Wei, An-Chi Tien, Wen-Shien Huang, Yuan-Chii G Lee, Chin-Lin Peng, Huei-Hun Tseng, Cheng-Yan Kao, and Chi-Ying F Huang. 2004. Point: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics* 20(17):3273–3276.

Hunter, Sarah, Rolf Apweiler, Teresa K. Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, Robert D. Finn, Julian Gough, Daniel Haft, Nicolas Hulo, Daniel Kahn, Elizabeth Kelly, Aurelie Laugraud, Ivica Letunic, David Lonsdale, Rodrigo Lopez, Martin Madera, John Maslen, Craig McAnulla, Jennifer McDowall, Jaina Mistry, Alex

Mitchell, Nicola Mulder, Darren Natale, Christine Orengo, Antony F. Quinn, Jeremy D. Selengut, Christian J. A. Sigrist, Manjula Thimma, Paul D. Thomas, Franck Valentin, Derek Wilson, Cathy H. Wu, and Corin Yeats. 2009. Interpro: the integrative protein signature database. *Nucl. Acids Res.* 37(suppl 1):D211–215.

Irizarry, Rafael A., Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostat* 4(2):249–264.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 98(8):4569–4574.

Jameson, David M, John C Croney, and Pierre D J Moens. 2003. Fluorescence: basic concepts, practical aspects, and some anecdotes. *Methods Enzymol* 360: 1–43.

Jang, Hyunchul, Jaesoo Lim, Joon-Ho Lim, Soo-Jun Park, Kyu-Chul Lee, and Seon-Hee Park. 2006. Finding the evidence for protein-protein interactions from pubmed abstracts. *Bioinformatics* 22(14):e220–e226.

Jansen, Ronald, and Mark Gerstein. 2004. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Current Opinion in Microbiology* 7(5):535–545.

Jansen, Ronald, Dov Greenbaum, and Mark Gerstein. 2002. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* 12(1):37–46.

Jansen, Ronald, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J. Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F. Greenblatt, and Mark Gerstein. 2003. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302(5644):449–453.

Jefferson, Emily R., Thomas P. Walsh, Timothy J. Roberts, and Geoffrey J. Barton. 2007. Snappi-db: a database and api of structures, interfaces and alignments for protein-protein interactions. *Nucl. Acids Res.* 35(suppl 1):D580–589.

Jensen, Lars J., Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, Peer Bork, and Christian von Mering. 2008. String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucl. Acids Res.* gkn760.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabsi. 2000. The large-scale organization of metabolic networks. *Nature* 407(6804):651–654.

Jiang, JJ, and DW Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *International conference research on computational linguistics (rocling x)*, 9008.

Joshi-Tope, G., M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 33(Database issue):D428–D432.

Jost, J. P., O. Munch, and T. Andersson. 1991. Study of protein-dna interactions by surface plasmon resonance (real time kinetics). *Nucleic Acids Res* 19(10):2788.

Kaminuma, Eli, Takehide Kosuge, Yuichi Kodama, Hideo Aono, Jun Mashima, Takashi Gojobori, Hideaki Sugawara, Osamu Ogasawara, Toshihisa Takagi, Kousaku Okubo, and Yasukazu Nakamura. 2011. Ddbj progress report. *Nucleic Acids Res* 39(Database issue):D22–D27.

Kanehisa, Minoru. 2002. The kegg database. *Novartis Foundation Symposium* 247: 91–101; discussion 101–103, 119–128, 244–252.

Kanehisa, Minoru, Susumu Goto, Masahiro Hattori, Kiyoko F Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. 2006. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Research* 34(Database issue):D354–357.

Kanehisa, Minoru, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. 2004. The kegg resource for deciphering the genome. *Nucleic Acids Res* 32(Database issue):D277–D280.

Karatzoglou, Alexandros, Alexandros Smola, Kurt Hornik, and Achim Zeileis. 2004. kernlab - an s4 package for kernel methods in r. *Journal of Statistical Software* 11(9):1–20.

Kendall, MG. 1938. A new measure of rank correlation. *Biometrika* 30(1/2):93, 81.

Kerrien, S., Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. 2007a. Intact–open source resource for molecular interaction data. *Nucl. Acids Res.* 35(suppl 1):D561–565.

Kerrien, Samuel, Sandra Orchard, Luisa Montecchi-Palazzi, Bruno Aranda, Antony Quinn, Nisha Vinod, Gary Bader, Ioannis Xenarios, Jerome Wojcik, David Sherman, Mike Tyers, John Salama, Susan Moore, Arnaud Ceol, Andrew Chatraryamontri, Matthias Oesterheld, Volker Stumpflen, Lukasz Salwinski, Jason Nerothin, Ethan Cerami, Michael Cusick, Marc Vidal, Michael Gilson, John Armstrong, Peter Woollard, Christopher Hogue, David Eisenberg, Gianni Cesareni, Rolf Apweiler, and Henning Hermjakob. 2007b. Broadening the horizon - level 2.5 of the hupo-psi format for molecular interactions. *BMC Biology* 5(1):44.

Kersey, Paul J, Jorge Duarte, Allyson Williams, Youla Karavidopoulou, Ewan Birney, and Rolf Apweiler. 2004. The international protein index: an integrated database for proteomics experiments. *Proteomics* 4(7):1985–1988.

Keshava Prasad, T. S., Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S. Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C. J. Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K. Kashyap, Riaz Mohmood, Y. L. Ramachandra, V. Krishna, B. Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. 2009. Human protein reference database2009 update. *Nucleic Acids Research* 37(Database issue):D767–D772.

Kim, Sun, Soo-Yong Shin, In-Hee Lee, Soo-Jin Kim, Ram Sriram, and Byoung-Tak Zhang. 2008. Pie: an online prediction system for protein-protein interactions from text. *Nucleic Acids Res* 36(Web Server issue):W411–W415.

King, A. D., N. Przulj, and I. Jurisica. 2004. Protein complex prediction via cost-based clustering. *Bioinformatics* 20(17):3013–3020.

King, R. D., and M. J. Sternberg. 1996. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Science : A Publication of the Protein Society* 5(11):2298–2310.

Kortemme, Tanja, and David Baker. 2002. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* 99(22): 14116–14121.

Lander, Gabriel C, Liang Tang, Sherwood R Casjens, Eddie B Gilcrease, Peter Prevelige, Anton Poliakov, Clinton S Potter, Bridget Carragher, and John E Johnson. 2006. The structure of an infectious p22 virion shows the signal for headful dna packaging. *Science* 312(5781):1791–1795.

Laplace, Pierre Simon. 1774. Memoir on the probability of the causes of events. *Statistical Science* 1(3):364–378.

Lee, S. J. 1991. Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure. *Proc Natl Acad Sci U S A* 88(10):4250–4254.

Lee, Sheng-An, Cheng hsiung Chan, Chi-Hung Tsai, Jin-Mei Lai, Feng-Sheng Wang, Cheng-Yan Kao, and Chi-Ying F Huang. 2008. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics* 9 Suppl 12:S11.

Lehner, Ben, and Andrew Fraser. 2004. A first-draft human protein-interaction map. *Genome Biology* 5(9):R63.

Lensink, Marc F, Ral Mndez, and Shoshana J Wodak. 2007. Docking and scoring protein complexes: Capri 3rd edition. *Proteins* 69(4):704–718.

Li, Cheng, and Wing Hung Wong. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology* 2(8):research0032.1–research0032.11.

Li, Cheng, and Wing Hung Wong. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America* 98(1):31–36.

Li, S. 2004. A map of the interactome network of the metazoan c. elegans. *Science* 303(5657):540–543.

Lim, Janghoo, Tong Hao, Chad Shaw, Akash J Patel, Gbor Szab, Jean-Franois Rual, C. Joseph Fisk, Ning Li, Alex Smolyar, David E Hill, Albert-Lszl Barabsi, Marc Vidal, and Huda Y Zoghbi. 2006. A protein-protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. *Cell* 125(4): 801–814.

Lin, Dekang. 1998. An information-theoretic definition of similarity. In *Proceedings of the fifteenth international conference on machine learning*, 296–304. Morgan Kaufmann Publishers Inc.

Lin, Mingzhi, Xueling Shen, and Xin Chen. 2010. Pair: the predicted arabidopsis interactome resource. *Nucleic Acids Res.*

Lindsay, M E, J M Holaska, K Welch, B M Paschal, and I G Macara. 2001. Ran-binding protein 3 is a cofactor for crm1-mediated nuclear protein export. *The Journal of Cell Biology* 153(7):1391–1402.

Lise, Stefano, Daniel Buchan, Massimiliano Pontil, and David T Jones. 2011. Predictions of hot spot residues at protein-protein interfaces using support vector machines. *PLoS One* 6(2):e16774.

Little, Roderick J.A., and Donald B. Rubin. 1987. *Statistical analysis with missing data*. Wiley John & Sons.

Liu, Guimei, Limsoon Wong, and Hon Nian Chua. 2009. Complex discovery from weighted ppi networks. *Bioinformatics* 25(15):1891–1897.

Liu, Sunbin, Reinhard Rauhut, Hans-Peter Vornlocher, and Reinhard Lhrmann. 2006. The network of proteinprotein interactions within the human u4/u6.u5 tri-snrnp. *RNA* 12(7):1418–1430.

Lofas, Stefan, and Bo Johnsson. 1990. A novel hydrogel matrix on gold surfaces in surface plasmon resonance sensors for fast and efficient covalent immobilization of ligands. *Journal of the Chemical Society, Chemical Communications* (21):1526–1528.

Lonhienne, Thierry G, Jade K Forwood, Mary Marfori, Gautier Robin, Bostjan Kobe, and Bernard J Carroll. 2009. Importin- is a gdp-to-gtp exchange factor of ran: Implications for the mechanism of nuclear import. *The J* 284(34):22549–22558.

Lord, P W, R D Stevens, A Brass, and C A Goble. 2003. Semantic similarity measures as tools for exploring the gene ontology. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 601–612.

MacBeath, G., and S. L. Schreiber. 2000. Printing proteins as microarrays for high-throughput function determination. *Science* 289(5485):1760–1763.

Marcotte, E. M., I. Xenarios, and D. Eisenberg. 2001. Mining literature for protein-protein interactions. *Bioinformatics* 17(4):359–363.

Marcotte, Edward M., Matteo Pellegrini, Ho-Leung Ng, Danny W. Rice, Todd O. Yeates, and David Eisenberg. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science* 285(5428):751–753.

Martin, Shawn, Diana Roe, and Jean-Loup Faulon. 2005. Predicting protein-protein interactions using signature products. *Bioinformatics* 21(2):218–226.

Matsumoto, Makoto, and Takuji Nishimura. 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.* 8(1):3–30.

Matthews, L. R., P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M. Vidal. 2001. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* 11(12):2120–2126.

Matthews, Lisa, Gopal Gopinath, Marc Gillespie, Michael Caudy, David Croft, Bernard de Bono, Phani Garapati, Jill Hemish, Henning Hermjakob, Bijay Jassal, Alex Kanapin, Suzanna Lewis, Shahana Mahajan, Bruce May, Esther Schmidt, Imre Vastrik, Guanming Wu, Ewan Birney, Lincoln Stein, and Peter D'Eustachio. 2009. Reactome knowledgebase of human biological pathways and processes. *Nucl. Acids Res.* 37(suppl 1):D619–622.

McDowall, Mark D., Michelle S. Scott, and Geoffrey J. Barton. 2009. Pips: human protein-protein interaction prediction database. *Nucl. Acids Res.* 37(suppl 1): D651–656.

von Mering, Christian, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. 2003. String: a database of predicted functional associations between proteins. *Nucleic Acids Res* 31(1):258–261.

von Mering, Christian, Lars J Jensen, Michael Kuhn, Samuel Chaffron, Tobias Doerks, Beate Krger, Berend Snel, and Peer Bork. 2007. String 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 35(Database issue):D358–D362.

von Mering, Christian, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A Huynen, and Peer Bork. 2005. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* 33(Database issue):D433–D437.

Mishra, Gopa R., M. Suresh, K. Kumaran, N. Kannabiran, Shubha Suresh, P. Bala, K. Shivakumar, N. Anuradha, Raghunath Reddy, T. Madhan Raghavan, Shalini Menon, G. Hanumanthu, Malvika Gupta, Sapna Upendran, Shweta Gupta, M. Mahesh, Bincy Jacob, Pinky Mathew, Pritam Chatterjee, K. S. Arun, Salil

Sharma, K. N. Chandrika, Nandan Deshpande, Kshitish Palvankar, R. Raghavnath, R. Krishnakanth, Hiren Karathia, B. Rekha, Rashmi Nayak, G. Vishnupriya, H. G. Mohan Kumar, M. Nagini, G. S. Sameer Kumar, Rojan Jose, P. Deepthi, S. Sujatha Mohan, T. K. B. Gandhi, H. C. Harsha, Krishna S. Deshpande, Malabika Sarker, T. S. Keshava Prasad, and Akhilesh Pandey. 2006. Human protein reference database–2006 update. *Nucl. Acids Res.* 34(suppl 1):D411–414.

Mitchell, Tom M. 1997. *Machine learning*. McGraw-Hill.

Mosteller, Frederick, and John W. Tukey. 1977. *Data analysis and regression: A second course in statistics*. 1st ed. Addison Wesley.

Mulder, Nicola J, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Virginie Buillard, Lorenzo Cerutti, Richard Copley, Emmanuel Courcelle, Ujjwal Das, Louise Daugherty, Mark Dibley, Robert Finn, Wolfgang Fleischmann, Julian Gough, Daniel Haft, Nicolas Hulo, Sarah Hunter, Daniel Kahn, Alexander Kanapin, Anish Kejariwal, Alberto Labarga, Petra S Langendijk-Genevaux, David Lonsdale, Rodrigo Lopez, Ivica Letunic, Martin Madera, John Maslen, Craig McAnulla, Jennifer McDowall, Jaina Mistry, Alex Mitchell, Anastasia N Nikolskaya, Sandra Orchard, Christine Orengo, Robert Petryszak, Jeremy D Selengut, Christian J A Sigrist, Paul D Thomas, Franck Valentin, Derek Wilson, Cathy H Wu, and Corin Yeats. 2007. New developments in the interpro database. *Nucleic Acids Research* 35(Database issue): D224–228.

Müller, Joachim D, Yan Chen, and Enrico Gratton. 2003. Fluorescence correlation spectroscopy. *Methods Enzymol* 361:69–92.

Ofran, Yanay, and Burkhard Rost. 2007a. Isis: interaction sites identified from sequence. *Bioinformatics* 23(2):e13–e16.

———. 2007b. Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol* 3(7):e119.

Pagel, Philipp, Stefan Kovac, Matthias Oesterheld, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Pekka Mark, Volker Stumpflen, Hans-Werner Mewes, Andreas Ruepp, and Dmitrij Frishman. 2005. The mips mammalian protein-protein interaction database. *Bioinformatics* 21(6): 832–834.

Parkinson, Helen, Misha Kapushesky, Nikolay Kolesnikov, Gabriella Rustici, Mohammad Shojatalab, Niran Abeygunawardena, Hugo Berube, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Ele Holloway, Margus Lukk, James Malone, Roby Mani, Ekaterina Pilicheva, Tim F. Rayner, Faisal Rezwan, Anjan Sharma, Eleanor Williams, Xiangqun Zheng Bradley, Tomasz Adamusiak, Marco Brandizi, Tony Burdett, Richard Coulson, Maria Krestyaninova, Pavel Kurnosov, Eamonn Maguire, Sudeshna Guha Neogi, Philippe Rocca-Serra, Susanna-Assunta Sansone,

Nataliya Sklyar, Mengyao Zhao, Ugis Sarkans, and Alvis Brazma. 2009. Arrayexpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucl. Acids Res.* 37(suppl 1):D868–872.

Pearson, Karl. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58:240–242.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* 96(8):4285–4288.

Peng, Kang, Predrag Radivojac, Slobodan Vucetic, A Keith Dunker, and Zoran Obradovic. 2006. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 7(1):208.

Peri, Suraj, J. Daniel Navarro, Ramars Amanchy, Troels Z Kristiansen, Chandra Kiran Jonnalagadda, Vineeth Surendranath, Vidya Niranjan, Babylakshmi Muthusamy, T. K B Gandhi, Mads Gronborg, Nieves Ibarrola, Nandan Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Zhixing Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Shubha Suresh, Neelanjana Ghosh, R. Saravana, Sreenath Chandran, Subhalakshmi Krishna, Mary Joy, Sanjeev K Anand, V. Madavan, Ansamma Joseph, Guang W Wong, William P Schiemann, Stefan N Constantinescu, Lily Huang, Roya Khosravi-Far, Hanno Steen, Muneesh Tewari, Saghi Ghaffari, Gerard C Blobe, Chi V Dang, Joe G N Garcia, Jonathan Pevsner, Ole N Jensen, Peter Roepstorff, Krishna S Deshpande, Arul M Chinnaiyan, Ada Hamosh, Aravinda Chakravarti, and Akhilesh Pandey. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 13(10):2363–2371.

Peri, Suraj, J. Daniel Navarro, Troels Z. Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, T. K. B. Gandhi, K. N. Chandrika, Nandan Deshpande, Shubha Suresh, B. P. Rashmi, K. Shanker, N. Padma, Vidya Niranjan, H. C. Harsha, Naveen Talreja, B. M. Vrushabendra, M. A. Ramya, A. J. Yatish, Mary Joy, H. N. Shivashankar, M. P. Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Neelanjana Ghosh, R. Saravana, Sreenath Chandran, Sujatha Mohan, Chandra Kiran Jonnalagadda, C. K. Prasad, Chandan Kumar-Sinha, Krishna S. Deshpande, and Akhilesh Pandey. 2004. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research* 32(Database issue):D497–D501.

Picard, Richard R., and R. Dennis Cook. 1984. Cross-validation of regression models. *Journal of the American Statistical Association* 79(387):575–583.

Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. 2007. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res.* 35(suppl 1):D61–65.

Puig, O., F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Sraphin. 2001. The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods* 24(3):218–229.

Qi, Yanjun, Ziv Bar-Joseph, and Judith Klein-Seetharaman. 2006. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* 63(3):490–500.

R Development Core Team. 2009. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ramani, Arun K, Razvan C Bunescu, Raymond J Mooney, and Edward M Marcotte. 2005. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology* 6(5):R40.

Ramani, Arun K, Zhihua Li, G Traver Hart, Mark W Carlson, Daniel R Boutz, and Edward M Marcotte. 2008. A map of human protein interactions derived from co-expression of human mrnas and their orthologs. *Mol Syst Biol* 4.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabasi. 2002. Hierarchical organization of modularity in metabolic networks. *Science* 297(5586): 1551–1555.

Reid, Adam J, Juan A G Ranea, Andrew B Clegg, and Christine A Orengo. 2010. Coda: accurate detection of functional associations between proteins in eukaryotic genomes using domain fusion. *PLoS One* 5(6):e10908.

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on artificial intelligence - volume 1*, 448–453. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.

Rhodes, Daniel R, Scott A Tomlins, Sooryanarayana Varambally, Vasudeva Mahavisno, Terrence Barrette, Shanker Kalyana-Sundaram, Debashis Ghosh, Akhilesh Pandey, and Arul M Chinnaiyan. 2005. Probabilistic model of the human protein-protein interaction network. *Nat Biotech* 23(8):951–959.

Rigaut, G., A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Sraphin. 1999. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* 17(10):1030–1032.

Rivera, Corban G, Rachit Vakil, and Joel S Bader. 2010. Nemo: Network module identification in cytoscape. *BMC Bioinformatics* 11 Suppl 1:S61.

Robinson, Carol V, Andrej Sali, and Wolfgang Baumeister. 2007. The molecular sociology of the cell. *Nature* 450(7172):973–982.

Rodgers, Joseph Lee, and W. Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician* 42(1):59–66.

Rossmann, Michael G., Marc C. Morais, Petr G. Leiman, and Wei Zhang. 2005. Combining x-ray crystallography and electron microscopy. *Structure* 13(3):355–362.

Rost, B., J. Liu, R. Nair, K. O. Wrzeszczynski, and Y. Ofran. 2003. Automatic prediction of protein function. *Cell Mol Life Sci* 60(12):2637–2650.

Rost, B, and C Sander. 1993. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology* 232(2):584–599.

Roth, Richard B, Peter Hevezi, Jerry Lee, Dorian Willhite, Sandra M Lechner, Alan C Foster, and Albert Zlotnik. 2006. Gene expression analyses reveal molecular relationships among 20 regions of the human cns. *Neurogenetics* 7(2):67–80.

Roy, Sushmita, Diego Martinez, Harriett Platero, Terran Lane, and Margaret Werner-Washburne. 2009. Exploiting amino acid composition for predicting protein-protein interactions. *PLoS ONE* 4(11):e7813.

Rual, Jean-Francois, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amelie Dricot, Ning Li, Gabriel F. Berriz, Francis D. Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S. Goldberg, Lan V. Zhang, Sharyl L. Wong, Giovanni Franklin, Siming Li, Joanna S. Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S. Sikorski, Jean Vandenhaute, Huda Y. Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E. Cusick, David E. Hill, Frederick P. Roth, and Marc Vidal. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437(7062):1173–1178.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature* 323:3533–536.

Salamov, A A, and V V Solovyev. 1995. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of Molecular Biology* 247(1):11–15.

Salwinski, Lukasz, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. 2004. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 32(Database issue):D449–D451.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science* 270(5235):467–470.

Schena, M., D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A* 93(20):10614–10619.

Schwarz, Gideon. 1978. Estimating dimension of a model. *ANNALS OF STATISTICS* 6(2):461–464.

Scott, Michelle, and Geoffrey Barton. 2007. Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinformatics* 8(1):239.

Shaughnessy, Patrick, Gary Livingston, and Michael V Graves. 2008. Towards predicting protein-protein interactions in novel organisms. *International Journal of Computational Biology and Drug Design* 1(3):235–253.

Shen, Juwen, Jian Zhang, Xiaomin Luo, Weiliang Zhu, Kunqian Yu, Kaixian Chen, Yixue Li, and Hualiang Jiang. 2007. Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences* 104(11):4337–4341.

Shoemaker, Benjamin A, and Anna R Panchenko. 2007. Deciphering proteinprotein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Computational Biology* 3(4):e43.

Shoyaib, Mohammad, M Abdullah-Al-Wadud, and Oksam Chae. 2009. Selecting negative examples for protein-protein interaction. *World Academy of Science, Engineering and Technology* 57:50–54.

Simons, A., N. Dafni, I. Dotan, Y. Oron, and D. Canaani. 2001a. Establishment of a chemical synthetic lethality screen in cultured human cells. *Genome Res* 11(2): 266–273.

Simons, A. H., N. Dafni, I. Dotan, Y. Oron, and D. Canaani. 2001b. Genetic synthetic lethality screen at the single gene level in cultured human cells. *Nucleic Acids Res* 29(20):E100.

Singh, Chingakham Ranjit, and Katsura Asano. 2007. Localization and characterization of protein-protein interaction sites. *Methods Enzymol* 429:139–161.

Sklar, Larry A, Mark B Carter, and Bruce S Edwards. 2007. Flow cytometry for drug discovery, receptor pharmacology and high-throughput screening. *Current Opinion in Pharmacology* 7(5):527–534.

Sleator, Roy D, and Paul Walsh. 2010. An overview of in silico protein function prediction. *Arch Microbiol* 192(3):151–155.

Smialowski, Pawel, Philipp Pagel, Philip Wong, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, Thomas Rattei, Dmitrij Frishman, and Andreas Ruepp. 2009. The negatome database: a reference set of non-interacting protein pairs. *Nucl. Acids Res.* gkp1026.

Smith, Graham R., and Michael J. E. Sternberg. 2002. Prediction of protein-protein interactions by docking methods. *Current Opinion in Structural Biology* 12(1): 28–35.

Smith, Mike, Victor Kunin, Leon Goldovsky, Anton J. Enright, and Christos A. Ouzounis. 2005. Magicmatch–cross-referencing sequence identifiers across databases. *Bioinformatics* 21(16):3429–3430.

Spearman, Charles Edward. 1904. The proof and measurement of association between two things. *American Journal of Psychology* 15:72–101.

Sperisen, Peter, and Marco Pagni. 2005. Jacop: A simple and robust method for the automated classification of protein sequences with modular architecture. *BMC Bioinformatics* 6(1):216.

Sprinzak, Einat, and Hanah Margalit. 2001. Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology* 311(4):681–692.

Sprinzak, Einat, Shmuel Sattath, and Hanah Margalit. 2003. How reliable are experimental protein-protein interaction data? *J Mol Biol* 327(5):919–923.

Stark, Chris, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. Biogrid: a general repository for interaction datasets. *Nucl. Acids Res.* 34(suppl 1):D535–539.

Stelzl, Ulrich, and Erich E Wanker. 2006. The value of high quality protein-protein interaction networks for systems biology. *Curr Opin Chem Biol* 10(6):551–558.

Stelzl, Ulrich, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H. Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, Jan Timm, Sascha Mintzlaff, Claudia Abraham, Nicole Bock, Silvia Kietzmann, Astrid Goedde, Engin Toks, Anja Droege, Sylvia Krobitsch, Bernhard Korn, Walter Birchmeier, Hans Lehrach, and Erich E. Wanker. 2005. A human protein-protein interaction network: A resource for annotating the proteome. *Cell* 122(6):957–968.

Stigler, Stephen M. 1986a. *The history of statistics: The measure of uncertainty before 1900*. Belknap Harvard.

———. 1986b. Laplace's 1774 memoir on inverse probability. *Statistical Science* 1(3):359–363.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* 36(2):111–147.

Stumpf, Michael P. H., Thomas Thorne, Eric de Silva, Ronald Stewart, Hyeong Jun An, Michael Lappe, and Carsten Wiuf. 2008. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences* 105(19):6959–6964.

Stumpf, Michael P.H., William P. Kelly, Thomas Thorne, and Carsten Wiuf. 2007. Evolution at the system level: the natural history of protein interaction networks. *Trends in Ecology & Evolution* 22(7):366–373.

Su, Andrew I., Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A. Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P. Cooke, John R. Walker, and John B. Hogenesch. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* 101(16):6062–6067.

Su, Chong, Jose M Peregrin-Alvarez, Gareth Butland, Sadhna Phanse, Vincent Fong, Andrew Emili, and John Parkinson. 2008. Bacteriome.org–an integrated protein interaction database for e. coli. *Nucleic Acids Res* 36(Database issue): D632–D636.

Swets, J A, R M Dawes, and J Monahan. 2000. Better decisions through science. *Scientific American* 283(4):82–87.

Tanay, Amos, Roded Sharan, and Ron Shamir. 2002. Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 18 Suppl 1:S136–S144.

Tompa, Peter, and Monika Fuxreiter. 2008. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends in Biochemical Sciences* 33(1):2–8.

Tong, A. H., M. Evangelista, A. B. Parsons, H. Xu, G. D. Bader, N. Pag, M. Robinson, S. Raghibizadeh, C. W. Hogue, H. Bussey, B. Andrews, M. Tyers, and C. Boone. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294(5550):2364–2368.

Tucker, Chandra L, and Stanley Fields. 2003. Lethal combinations. *Nat Genet* 35(3):204–205.

Tulipano, Angelica, Giacinto Donvito, Flavio Licciulli, Giorgio Maggi, and Andreas Gisel. 2007. Gene analogue finder: a grid solution for finding functionally analogous gene products. *BMC Bioinformatics* 8(1):329.

Uetz, P., L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. 2000. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature* 403(6770):623–627.

Valet, G. 2005. Cytomics, the human cytome project and systems biology: top-down resolution of the molecular biocomplexity of organisms by single cell analysis. *Cell Proliferation* 38(4):171–174.

Vallabhajosyula, Ravishankar R., Deboki Chakravarti, Samina Lutfeali, Animesh Ray, and Alpan Raval. 2009. Identifying hubs in protein interaction networks. *PLoS ONE* 4(4):e5344.

Vastrik, Imre, Peter D'Eustachio, Esther Schmidt, Geeta Joshi-Tope, Gopal Gopinath, David Croft, Bernard de Bono, Marc Gillespie, Bijay Jassal, Suzanna Lewis, Lisa Matthews, Guanming Wu, Ewan Birney, and Lincoln Stein. 2007. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 8(3):R39.

Venkatesan, Kavitha, Jean-Francois Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, Muhammed A Yildirim, Nicolas Simonis, Kathrin Heinzmann, Fana Gebreab, Julie M Sahalie, Sebiha Cevik, Christophe Simon, Anne-Sophie

de Smet, Elizabeth Dann, Alex Smolyar, Arunachalam Vinayagam, Haiyuan Yu, David Szeto, Heather Borick, Amelie Dricot, Niels Klitgord, Ryan R Murray, Chenwei Lin, Maciej Lalowski, Jan Timm, Kirstin Rau, Charles Boone, Pascal Braun, Michael E Cusick, Frederick P Roth, David E Hill, Jan Tavernier, Erich E Wanker, Albert-Laszlo Barabasi, and Marc Vidal. 2009. An empirical framework for binary interactome mapping. *Nat Meth* 6(1):83–90.

Wagner, A., and D. A. Fell. 2001. The small world inside large metabolic networks. *Proc Biol Sci* 268(1478):1803–1810.

Walhout, A. J., R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg, and M. Vidal. 2000. Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science* 287(5450):116–122.

Wang, Bing, Peng Chen, De-Shuang Huang, Jing jing Li, Tat-Ming Lok, and Michael R Lyu. 2006. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett* 580(2):380–384.

Wang, Bing, Peng Chen, Peizhen Wang, Guangxin Zhao, and Xiang Zhang. 2010. Radial basis function neural network ensemble for predicting protein-protein interaction sites in heterocomplexes. *Protein Pept Lett* 17(9):1111–1116.

Wheeler, David L, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael Dicuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, Oleg Khovayko, David Landsman, David J Lipman, Thomas L Madden, Donna R Maglott, Vadim Miller, James Ostell, Kim D Pruitt, Gregory D Schuler, Martin Shumway, Edwin Sequeira, Steven T Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L Tatusov, Tatiana A Tatusova, Lukas Wagner, and Eugene Yaschenko. 2008. Database resources of the national center for biotechnology information. *Nucleic Acids Research* 36(Database issue):D13–21.

Wiles, Amy, Mark Doderer, Jianhua Ruan, Ting-Ting Gu, Dashnamoorthy Ravi, Barron Blackman, and Alexander Bishop. 2010. Building and analyzing protein interactome networks by cross-species comparisons. *BMC Systems Biology* 4(1): 36.

Xiao, Chuan, and Michael G Rossmann. 2007. Interpretation of electron density with stereographic roadmap projections. *J Struct Biol* 158(2):182–187.

Yan, Yuling, and Gerard Marriott. 2003. Analysis of protein interactions using fluorescence technologies. *Curr Opin Chem Biol* 7(5):635–640.

Yang, Yong, Hong Wang, and Dorothy A Erie. 2003. Quantitative characterization of biomolecular assemblies and interactions using atomic force microscopy. *Methods* 29(2):175–187.

Ye, P, B D Peyser, X Pan, J D Boeker, F A Spencer, and J S Bader. 2005. Gene function prediction from congruent synthetic lethal interactions in yeast. *Molecular Systems Biology* 26:1–12.

Yook, Soon-Hyung, Zoltn N. Oltvai, and Albert-Lszl Barabsi. 2004. Functional and topological characterization of protein interaction networks. *PROTEOMICS* 4(4): 928–942.

Zhang, C T, and R Zhang. 2001. A refined accuracy index to evaluate algorithms of protein secondary structure prediction. *Proteins* 43(4):520–522.

Zhang, Chun-Ting, and Ren Zhang. 2003. Q9, a content-balancing accuracy index to evaluate algorithms of protein secondary structure prediction. *The International Journal of Biochemistry & Cell Biology* 35(8):1256–1262.

Zhang, Ying, Boguslaw Stec, and Adam Godzik. 2007. Between order and disorder in protein structures: analysis of "dual personality" fragments in proteins. *Structure* 15(9):1141–1147.

Zhou, H. X., and Y. Shan. 2001. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* 44(3):336–343.

Zhu, H., M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. A. Dean, M. Gerstein, and M. Snyder. 2001. Global analysis of protein activities using proteome chips. *Science* 293(5537):2101–2105.

Zvelebil, M J, G J Barton, W R Taylor, and M J Sternberg. 1987. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *Journal of Molecular Biology* 195(4):957–961.

. . . she said yes.